Fairness of Artificial Intelligence in Predicting Recidivism Risk and Beyond.

Michael Mayowa Farayola, B. Tech, M. Tech

Supervised by Irina Tal, Regina Connolly, Malika Bendechache (University of Galway)

A thesis presented for the degree of Doctor of Philosophy

School of Computing

Dublin City University

December, 2025

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work, and that I have exercised reasonable care to ensure that the work is original and have conformed to the regulations on the use and declaration of Generative AI, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: Michael Mayowa Farayola

ID No: 22262402

Date: 31 December 2025

Appendix A – Reporting of Generative AI Tool Use

Tools

- **ChatGPT (OpenAI GPT-5)** — used between August 2025 and October 2025 for improving the readability and clarity of academic writing during thesis drafting.
- **Grammarly Premium** — used during the same period for grammar correction, punctuation, and style consistency.

Outputs

ChatGPT was used to suggest re-phrasings and simplify complex academic passages. Grammarly was employed to proof-read and refine sentence structure and ensure linguistic accuracy before final submission.

Prompts

Typical prompts included requests such as "Rephrase this paragraph for academic clarity." or "Suggest a more concise and formal alternative."

Iterations

Several iterations were conducted to refine sentence flow and maintain authorial tone. All AI outputs were critically reviewed and verified by the author and rewritten where necessary.

Critical Reflection

I found the use of Generative AI tools helpful in improving grammar and text

readability, which is important as I am not a native English speaker. Their use was limited to language refinement and did not contribute to research design, analysis, or results.

Acknowledgements

I wish to convey my profound appreciation to my supervisors, **Dr. Irina Tal**, **Dr. Takfarinas Saber**, **Prof. Regina Connolly**, and **Dr. Malika Bendechache**, for their invaluable guidance, support, and encouragement throughout this research. Their expertise, insights, and constructive feedback have been instrumental in shaping this thesis, and I am deeply grateful for the time and effort they have dedicated to my academic journey.

I also extend my sincere thanks to **Dublin City University** and **LERO – the Science Foundation Ireland Research Centre for Software (SFI)** for providing the resources and opportunities that made this research possible.

My deepest gratitude goes to my **dad, mum, and siblings**, whose unwavering love, support, and patience have been a constant source of strength. Their belief in me has sustained my motivation and helped me overcome challenges along the way, making this achievement possible.

Finally, I would like to thank my extended family and friends for their understanding, encouragement, and support. I am sincerely grateful to all who have contributed, directly or indirectly, to the completion of this thesis. This accomplishment would not have been possible without your collective support.

Publications

The following publications have resulted from the research presented in this thesis:

1. Michael Mayowa Farayola, Irina Tal, Regina Connolly, et al. (2023). "Ethics and Trustworthiness of AI for Predicting the Risk of Recidivism: A Systematic Literature Review". In: *Information* 14.8, p. 426. DOI: https://doi.org/10.3390/info14080426

2. Michael Mayowa Farayola, Irina Tal, Bendechache Malika, et al. (2023). "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. DOI: https://doi.org/10.1145/3600160.3605033

3. Michael Mayowa Farayola, Malika Bendechache, Takfarinas Saber, et al. (2024b). "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

4. Michael Mayowa Farayola, Bendechache Malika, et al. (2026). "Beyond Calibration: Rethinking Algorithmic Fairness Through an Intersectional, Justice-Aware Lens". In

5. Michael Mayowa Farayola, Irina Tal, Takfarinas Saber, et al. (2025). "A fairness-focused approach to recidivism prediction: implications for accuracy, trust, and equity". In: *AI & SOCIETY*, pp. 1–19. DOI: https://doi.org/10.1007/s00146-025-02452-1

6. Michael Farayola et al. (2025). "Beyond Aggregate Fairness: Intersectional Auditing Across the AI Fairness Pipeline". In: *AI and Ethics*

7. Michael Mayowa Farayola, Malika Bendechache, Saber Takfarinas, et al. (2025). "Investigating Fairness-Aware Oversampling Strategies and Techniques Across Diverse Machine Learning Algorithms for Recidivism Prediction". In: *Minds and Machines* 35.3, p. 37

8. Michael Mayowa Farayola, Shane Kennedy, et al. (2025). "Intersectional Fairness in Healthcare AI: A Pipeline-Wide Evaluation of Multi-Stage Mitigation Strategies". In

Contents

List of Figures

List of Tables

Michael Mayowa Farayola

Abstract

This thesis designs, evaluates, and deploys fairness-aware artificial intelligence (AI) systems for recidivism prediction and other high-stakes domains. It addresses persistent gaps in the literature by integrating fairness interventions across the AI lifecycle, improving fairness–accuracy trade-offs, and developing methods that expose intersectional harms. The research advances multi-phase fairness pipelines that integrate pre-, in-, and post-processing techniques, supported by optimization-informed evaluation and intersectional auditing frameworks.

Grounded in trustworthy AI and socio-technical governance principles, the study aligns fairness interventions with ethical and procedural standards for responsible AI in criminal justice and beyond. Empirical validation uses four datasets: COMPAS, RisCanvi (a curated dataset), Adult Income, and an Irish health insurance dataset. Standardized protocols and fairness metrics, Statistical Parity Difference, Disparate Impact, Equal Opportunity Difference, and Predictive Equality Difference, guide analysis and model evaluation.

The thesis extends the AIF360 toolkit through enhanced algorithms (DIR+, AD+) capable of handling multi-valued protected attributes and improving training stability. Results show that integrated pipelines outperform isolated methods, achieving consistent fairness gains with limited accuracy reduction. Fairness-aware oversampling mitigates subgroup imbalance but requires careful validation to avoid

synthetic bias. Intersectional auditing reveals disparities masked in aggregate measures and highlights the contextual nature of fairness.

Conceptually, the thesis contributes the Justice-Aware AI Fairness (JAAF) framework, linking technical fairness methods with intersectional and governance-based accountability. Overall, it delivers integrated fairness pipelines, optimization-based evaluation tools, and guidelines for fair, transparent, and trustworthy AI systems in high-stakes decision-making.

Chapter 1

Introduction

1.1 Background and Context

The integration of Artificial Intelligence (AI) into criminal justice systems has transformed how the institutions make decisions about crime, punishment, and rehabilitation. This transformation centers on the pursuit of more consistent, data-driven methods that assess offender risk, particularly the risk of recidivism, the likelihood that a convicted individual will re-offend after release (Sushina and Sobenin 2020; Engel, Linhardt, and Schubert 2025; Christian 2024). Researchers and practitioners refer to these computational systems as *risk assessment tools* (J. Joseph 2025; Christian 2024; Cavus et al. 2025), and thus play an increasingly central role in high-stakes decision-making across courts, correctional systems, and parole boards.

1.1.1 Historical Evolution of Risk Assessment

Risk assessments in the justice system predate the rise of modern AI and machine learning (ML) methods. Early approaches were based on *clinical judgment*, where parole officers, judges, or psychologists used professional experience to evaluate whether an individual posed a risk of re-offending (Bonta and S. C. Lee 2025). While such judgments reflected professional expertise, they were criticized for subjectivity, inconsistency, and susceptibility to individual biases (Desmarais, Johnson,

and Singh 2016; Viljoen et al. 2025). In the historical literature, this unstructured professional judgment is commonly described as the *first generation* of risk assessment, dominant through much of the twentieth century and characterized by informal, case-by-case evaluations (Bonta and Donald A Andrews 2007; Bonta and Donald Arthur Andrews 2023; Fazel et al. 2024)

A second wave of approaches introduced *actuarial methods* (Monahan and J. L. Skeem 2013; Bonta and Donald A Andrews 2007), statistical tools that relied on fixed factors such as age, prior convictions, or offense type to calculate risk scores. From the 1970s onward, jurisdictions began to adopt actuarial instruments that summed empirically associated risk factors into overall scores, enabling more consistent classification of offenders' likelihood of re-offending (Fazel et al. 2024). These methods improved upon purely clinical approaches by offering a systematic, replicable framework, and they demonstrated stronger predictive performance than unstructured professional judgment in comparative evaluations (Bonta and Donald A Andrews 2007; Viljoen et al. 2025). At the same time, actuarial approaches were often *atheoretical* and emphasized largely static, historical correlates (e.g., criminal history), limiting their interpretability and responsiveness to change over time (Cadigan and C. T. Lowenkamp 2011; Green 2020; Gundhus 2024).

Subsequent developments addressed these limitations by incorporating *dynamic* (changeable) risk factors and linking assessment to intervention. Building on the *risk-need-responsivity (RNR)* framework (Bonta and Donald A Andrews 2007; Fazel et al. 2024; Duwe and V. Clark 2025; Bonta and Donald Arthur Andrews 2023), the *third generation* of instruments (often termed "risk–need") retained static predictors but added structured measurement of criminogenic needs, such as antisocial associates, substance misuse, or unstable employment, so that scores could change as a person's circumstances changed, and so that supervision and treatment could target the drivers of risk (Bonta and Donald A Andrews 2007; Fazel et al. 2024). This evolution was followed by *fourth generation* tools that integrate case management functions (e.g., goal-setting, responsivity tailoring, ongoing monitoring)

with risk–need assessment, operationalizing RNR's full set of principles in day-to-day correctional practice (Bonta and Donald A Andrews 2007; Duwe and V. Clark 2025). The third- and fourth-generation advances established the modern view of risk assessment as both *predictive* and *prescriptive*: not merely sorting individuals by risk, but also informing what should be addressed and how (Bonta and Donald A Andrews 2007; Duwe and V. Clark 2025; Bonta and Donald Arthur Andrews 2023).

In parallel with these developments, a more recent wave coincided with advances in computational power and data availability, leading to the adoption of *AI- and ML-based algorithms*. These models are capable of analyzing large, complex datasets, uncovering non-linear patterns, and delivering predictions with high precision. They offer flexibility in incorporating a broader range of features, including demographic, socioeconomic, and behavioral data, thus appearing to promise more precise and equitable decision-making (R. Berk and Bleich 2014; Caroline Wang et al. 2022; Cavus et al. 2025; Christian 2024). While this AI/ML wave is distinct from the third- and fourth-generation RNR developments, it builds upon their insight that risk assessment must be systematic, evidence-based, and linked to intervention.

Applications, Adoption and Perceived Benefits

Today, agencies in the criminal justice system use AI-based risk assessment tools at various stages of the decision-making process. *Pre-trial tools* estimate the likelihood that an offender will fail to appear in court, whereas *post-conviction tools* focus on predicting long-term recidivism risk and play a particularly relevant role in parole and probation contexts (Cadigan and C. T. Lowenkamp 2011; Green 2020; U.S. Department of Justice 2024; Law Commission of Ontario 2025). This thesis focuses primarily on the post-conviction setting, where risk predictions influence decisions with life-altering consequences for individuals returning to society.

One of the most prominent tools in practice, the *Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)*, originates from Northpointe and operates across jurisdictions in the United States. Since its introduction, COM-

PAS has assessed more than one million offenders and continues to shape judicial and correctional decisions at scale (Lo Piano 2020; Engel, Linhardt, and Schubert 2025; Räz 2024; Green 2020). Its influence sparks intense debate regarding the appropriateness of entrusting algorithmic systems with decisions that determine liberty, punishment, and rehabilitation (Angwin et al. 2022; Levin n.d.; Chouldechova 2017; Dressel and Farid 2018; Rudin, Caroline Wang, and Coker 2020; McKay 2020).

The appeal of AI in criminal justice stems from a combination of long-standing system pressures, including case backlogs, prison overcrowding, and inconsistent decision-making, as well as escalating costs and a policy shift toward evidence-based practice (OECD 2025). Proponents argue that algorithmic risk assessment provides a way to standardize decisions, target scarce resources more effectively, and align supervision and treatment with empirically identified risk and need profiles (R. Berk and Bleich 2014; Desmarais, Johnson, and Singh 2016; Green 2018; Green 2020; Caroline Wang et al. 2022). Within the broader trustworthy-AI discourse, these proponents frame such claims as efforts that advance fairness, accountability, and societal well-being when systems operate under responsible design and governance (H. AI 2019; Thiebes, Lins, and Sunyaev 2021; Cofone and Khern-am-nuai 2024). Hence, the following points outline the key perceived benefits associated with the use of AI within the criminal justice system:

- **Enhanced efficiency and consistency.** By processing large, heterogeneous datasets at scale, AI-enabled tools deliver rapid, standardized assessments across cases and jurisdictions, reduce variability associated with unstructured professional judgment, and help courts and correctional centres manage high volumes criminal cases (OECD 2025; R. Berk and Bleich 2014; Desmarais, Johnson, and Singh 2016). Beyond speed, standardized scoring serves as a means to curb ad-hoc practices and promote more uniform application of policy guidelines (Green 2018; Law Commission of Ontario 2025).

- **Reduce unnecessary incarceration through risk-based triage.** Risk scores identify individuals suitable for non-custodial sanctions, diversion, or

less intensive supervision, thereby alleviating overcrowding without compromising supervision goals. This risk-based allocation supports proportionality and more humane sanctioning while preserving public safety (Green 2020; Caroline Wang et al. 2022).

- **Lower system costs via targeted interventions.** When supervision intensity and program enrollment match assessed risk and need, agencies concentrate higher-cost services on those most likely to benefit, improving outcomes at lower overall cost (R. Berk and Bleich 2014; Desmarais, Johnson, and Singh 2016; Fazel et al. 2024). This resource aligns with rehabilitation strategies that prioritize criminogenic needs and responsivity considerations (Bonta and S. C. Lee 2025; Jung et al. 2025).

- **Improve fairness by mitigating individual cognitive biases.** Advocates contend that structured, data-driven assessment tempers idiosyncratic human errors and implicit biases that influence judicial and parole decisions. When properly specified, validated, and monitored, models contribute to more equitable treatment across demographic groups and to measurable public-safety gains (Mohler and Porter 2021; H. AI 2019; R. A. Berk, Kuchibhotla, and Tchetgen Tchetgen 2024; Neil and Zanger-Tishler 2025).

- **Increase public safety through evidence-based supervision.** Risk estimates guide the intensity, timing, and type of supervision and treatment (e.g., substance-use programs, employment services), enabling agencies to intervene earlier and more precisely with higher-risk individuals while avoiding over-supervision of lower-risk cases (Green 2018; Green 2020; Bonta and Donald A Andrews 2007; Fazel et al. 2024).

- **Support transparency, auditability, and policy evaluation.** When agencies adopt documentation, validation, and monitoring practices, algorithmic assessments provide auditable rationales for decisions and facilitate longitudinal evaluation of policy impacts (e.g., pre-trial appearance, revocation, or

re-arrest outcomes) (H. AI 2019; Adler, Antoine, and Al-Saadoon 2024; Hurlburt 2017; Thiebes, Lins, and Sunyaev 2021; McCormack and Bendechache 2025).

Empirical studies report that well-specified risk models achieve predictive performance comparable to, and sometimes exceed, unaided human judgment in forecasting re-offending and related outcomes, thereby reinforcing their use as decision-support rather than decision-replacements (McKay 2020; Cavus et al. 2025; R. Berk and Bleich 2014). At the same time, practitioners realize these benefits in practice only when they develop models carefully, validate them rigorously across subpopulations, establish clear governance, and maintain ongoing monitoring.

Early Criticisms and Enduring Challenges

Before AI-based tools became prominent, scholars criticized earlier risk assessment instruments in ways that foreshadow current debates. They warned against the dangers of treating human behavior as a probabilistic output of statistical models (R. A. Berk, Kuchibhotla, and Tchetgen Tchetgen 2024; Wenzelburger, Yeung, and Hartmann 2025; Kristofik 2025). They highlighted the inability of such tools to capture the nuanced social and psychological realities that underlie criminal activity (Chugh 2021). Other researchers question whether risk assessment models, regardless of statistical sophistication, serve as ethically legitimate arbiters of justice (Alikhademi et al. 2021; UK Justice 2025; Law Commission of Ontario 2025; Montreal AI Ethics Institute 2025).

These early concerns intensify with the adoption of AI and ML, as the complexity of algorithms, the opacity of model decision processes, and the potential for embedding systemic biases amplify public and scholarly scrutiny(Kristofik 2025; Wenzelburger, Yeung, and Hartmann 2025). Consequently, while AI and ML introduce unprecedented opportunities to modernize criminal justice, they raise profound questions about *legitimacy, fairness, accountability, and trustworthiness*, questions that this thesis directly engages with in the context of recidivism prediction(UK

Justice 2025; Law Commission of Ontario 2025).

1.1.2 Risk Assessment and Recidivism Prediction

Recidivism refers to the tendency of a convicted individual to relapse into criminal behavior, often measured through re-arrest, re-conviction, or re-incarceration after release (Jain, Huber, Fegaras, et al. 2019; J. Zeng, Ustun, and Rudin 2017; Shih, Chiu, and Chou 2019; J. Skeem and C. Lowenkamp 2020). Recidivism poses a persistent challenge for criminal justice systems worldwide, as high rates of re-offending undermine public safety, strain correctional resources, and erode confidence in rehabilitation programs (Wijenayake, Graham, and Christen 2018; Shih, Chiu, and Chou 2019; Silva et al. 2025).

Evidence from multiple jurisdictions underscores the seriousness of recidivism. For example, a study by the Irish Prison Service, based on re-offending and re-conviction data from 2013, reported that 63.3% of released prisoners re-offended within three years, with more than 80% re-offending within twelve months of release (Donnellan 2013). The sample included 7,701 offenders, with a gender distribution heavily skewed toward males (92.5%). Men exhibited a higher recidivism rate (63%) compared to women (57%), yet the female rate still showed that more than half of female offenders re-offended. Comparable patterns have been documented in other contexts, such as the United States and Europe (Ozkan 2017; O'Donnell 2020), highlighting the widespread nature of the issue.

These statistics reveal not only the scale but also the complexity of recidivism, as the rates are influenced by demographic factors such as age and gender. Younger offenders tend to have higher recidivism rates, while socioeconomic disadvantage and limited access to rehabilitation services increase the likelihood of re-offending. As such, the effective assessment of recidivism risk remains critical for reducing prison overcrowding and designing interventions that enhance public safety.

Risk Assessment in Practice

Risk assessment tools serve as central instruments for anticipating recidivism and structuring decisions across the criminal justice lifecycle. In practice, institutions embed these tools at multiple procedural junctures and use them to allocate supervision intensity, tailor interventions, and document decision rationales (J. X. Han, Greenewald, and Shah 2025; Green 2018). Although practitioners also apply such tools at the pre-trial stage (e.g., failure-to-appear and new-arrest risk), this thesis focuses on post-conviction contexts, where assessments inform long-term supervision and rehabilitation (Cadigan and C. T. Lowenkamp 2011; Green 2020; California Department of Corrections and Rehabilitation 2025).

Operationally, criminal justice agencies use risk assessment instruments at multiple decision points to inform sentencing, supervision, and rehabilitation strategies. During sentencing and sanctioning, risk and need profiles guide decisions about the balance between custodial and community sanctions as well as the intensity of conditions attached to a sentence (Fazel et al. 2024; Green 2018). In parole and probation, risk scores influence release eligibility, supervision levels, and revocation responses, thereby aligning resource allocation with empirically assessed levels of risk (Desmarais, Johnson, and Singh 2016; Green 2020). For rehabilitation and reintegration, structured assessments of criminogenic needs support referrals to targeted interventions, such as substance use treatment or employment programs, and facilitate ongoing monitoring of progress over time (Bonta and Donald A Andrews 2007; Bonta and S. C. Lee 2025). Collectively, these applications illustrate how AI-driven risk instruments shape key operational decisions across the criminal justice lifecycle, linking predictive analytics to broader objectives of efficiency, proportionality, and offender rehabilitation.

What the tools compute. Most instruments combine static (historical, unchangeable) predictors such as age at first arrest or prior convictions with dynamic (changeable) factors such as employment, peers, substance use, or housing stability

(Karimi-Haghighi and Castillo 2021a; Karimi-Haghighi and Castillo 2021b; J. X. Han, Greenewald, and Shah 2025). These inputs produce scores that map to *risk bands* (e.g., low/medium/high) or probabilities, sometimes alongside need domains that highlight intervention targets (Fazel et al. 2024; Desmarais, Johnson, and Singh 2016). Agencies typically adopt cutoffs or decision guidelines that link risk bands to supervision intensity and program eligibility, with documentation requirements that standardize practice and support auditability (Desmarais, Johnson, and Singh 2016; H. AI 2019; European Parliament and Council 2024; N. AI 2023).

Workflow and governance. To embed tools within routine practice, agencies establish procedures for (i) *initial assessment* (at intake or sentencing), (ii) *periodic reassessment* to capture changes in dynamic needs, and (iii) *case planning* that links assessed needs to specific services and responsivity considerations (Bonta and Donald A Andrews 2007; K. G. Sheppard, Talaugon, and Hernandez 2024). Good practice emphasizes local validation, documentation of overrides (where practitioners depart from tool recommendations with reasons), and continuous quality assurance (calibration checks, subgroup analyses, drift monitoring) to ensure responsible and equitable use (Fazel et al. 2024; Hurlburt 2017; H. AI 2019). These governance steps ensure that tools function as decision supports rather than automated arbiters.

Examples in the field. Among widely used instruments, the *Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)* plays a prominent role in U.S. jurisdictions (Engel, Linhardt, and Schubert 2025). Drawing on criminal history, substance use, and socioeconomic indicators, COMPAS produces risk classifications that courts and parole boards use in a range of decisions. Its scale of adoption, affecting decisions for more than one million offenders, makes it a focal point for both operational practice and scholarly scrutiny (Lo Piano 2020; Green 2018; Green 2020). Beyond COMPAS, risk–need–responsivity (RNR)–oriented assessments align supervision and treatment with criminogenic needs, linking assessment outputs to program pathways and progress reviews (Bonta and Donald A

Andrews 2007).

Similarly, the Catalan prison system in Spain deploys the RisCanvi protocol (introduced in 2009) as a multi-level risk assessment administered periodically, typically every six months, and reviewed by a professional committee. RisCanvi comprises two versions: a screening tool (RisCanvi-S, 10 items) and a comprehensive version (RisCanvi-C, 43 items) that spans five domains, Criminal/Penitentiary, Biographical, Family/Social, Clinical, and Attitudes/Personality (Sánchez de Ribera et al. 2025). These domains encompass both static and dynamic risk factors and produce multiple outcome scores, including violent recidivism (REVI) and general recidivism (REGE) (Karimi-Haghighi and Castillo 2021b; Karimi-Haghighi and Castillo 2021a).

When researchers compared the hand-crafted RisCanvi scoring system with machine-learned models trained on historical Catalan data, the machine learning approaches achieved Area Under the Curve (AUC) scores of approximately 0.76 for violent recidivism and 0.73 for general recidivism, compared with 0.72 and 0.70, respectively, for the manual formula, representing modest but meaningful improvements in predictive performance (Karimi-Haghighi and Castillo 2021b). However, analyses also revealed disparities in prediction outcomes across nationality and age groups, underscoring persistent fairness concerns even within advanced data-driven frameworks (Karimi-Haghighi and Castillo 2021b).

A complementary line of work treated RisCanvi as a recurring assessment. By learning to predict which individuals are most likely to change risk before the next review, practitioners can selectively schedule evaluations, roughly halving the number of evaluations (to ∼50% while keeping missed risk changes around ∼14–15%. Fairness adjustments (equalizing evaluation rates across nationality and age) remain feasible with only small additional missed changes, preserving operational efficiency alongside equity goals (Karimi-Haghighi and Castillo 2021a).

Nevertheless, in day-to-day operations, agencies adopt risk tools to (i) standardize judgments across officers and jurisdictions, (ii) prioritize resources for higher-risk

cases, (iii) document decisions for oversight, and (iv) target interventions toward needs most associated with re-offending (Scaria et al. 2024; Green 2018; H. AI 2019). Realizing these benefits depends on careful model development and validation, staff training, clarity about permissible overrides, and monitoring for disparate impacts across demographic groups, especially where assessments inform liberty-affecting decisions (H. AI 2019; Green 2020).

Despite their widespread adoption, researchers and practitioners have repeatedly challenged the credibility and fairness of COMPAS and similar tools. In 2016, ProPublica (Angwin et al. 2016a) conducted a high-profile investigation into COMPAS and reported that the tool exhibited racial bias, producing false-positive predictions for Black offenders at nearly twice the rate of White offenders (Rodolfa et al. 2020). The report sparked intense debate on the role of AI in the criminal justice system and became a landmark example of how algorithmic tools can reproduce structural inequities.

Subsequent studies contested ProPublica's conclusions. Dieterich et al. (Dieterich, Mendoza, and Brennan 2016) and Flores et al. (Flores, Bechtel, and C. T. Lowenkamp 2016) argued that the analysis used flawed statistical measures and noted that, when predictive values were examined, COMPAS did not exhibit systematic racial bias in the way originally claimed. Nevertheless, critics pointed out that even when statistical definitions of fairness are satisfied, disparities in error rates across demographic groups remain ethically troubling.

Beyond race, researchers have also documented gender bias. Hamilton (Hamilton 2019) argued that COMPAS exhibited disparate impacts against women, raising questions about whether the tool adequately accounts for differences in recidivism patterns between men and women (Rief, R. A. Lewis, and Applegarth 2025). These concerns highlight the multidimensional nature of fairness, where aggregate performance measures may obscure subgroup harms.

Implications for Trust and Policy

The controversies surrounding COMPAS illustrate a broader dilemma for AI in criminal justice: while predictive tools hold promise for improving efficiency and consistency, they also risk eroding trust when they reproduce or exacerbate historical inequalities (Engel, Linhardt, and Schubert 2025; Khorshidi, Carter, and Mohler 2020; U.S. Department of Justice 2024). Stakeholders such as judges, policymakers, and civil rights advocates remain divided between those who view AI risk assessment tools as valuable decision-support mechanisms and those who caution against overreliance on opaque and potentially biased systems (Cavus et al. 2025).

At the policy level, these debates intersect with broader efforts to establish guidelines for *trustworthy AI*. For instance, the European Commission's Ethics Guidelines for Trustworthy AI (H. AI 2019) emphasize fairness as a central requirement for public trust in algorithmic decision-making. In the context of criminal justice, where predictions directly affect liberty and rehabilitation, ensuring fairness represents not only a technical challenge but also a societal and ethical imperative.

In summary, the prediction of recidivism risk illustrates both the potential and the pitfalls of applying AI to high-stakes domains. Courts and correctional agencies have widely deployed risk assessment tools like COMPAS to guide judicial decisions, yet their use has sparked controversies around fairness, bias, and legitimacy. These tensions highlight the importance of designing methods that enhance fairness across diverse populations, build trust among stakeholders, and align predictive systems with broader principles of justice. These considerations provide the foundation for the problem statement presented in the following subsection.

1.2 Problem Statement

The adoption of AI and machine learning models in criminal justice, particularly for recidivism risk prediction, introduces both opportunities and challenges. On the one hand, courts and correctional agencies have widely deployed AI risk assessment

tools such as COMPAS to support decisions about parole, probation, and sentencing, influencing the outcomes for millions of offenders (Desmarais, Johnson, and Singh 2016; Green 2020; Lo Piano 2020). These AI systems promise efficiency, consistency, and a reduction of subjective human bias. On the other hand, mounting evidence shows that such tools reinforce or even exacerbate existing inequities, particularly along sensitive attributes such as race and gender (Rodolfa et al. 2020; Hamilton 2019; Oatley 2022).

Despite extensive research into fairness in machine learning, several critical problems remain unresolved in the specific context of recidivism prediction and in general across domains:

1. **Fragmented fairness interventions.** Researchers often apply existing fairness-improving techniques in isolation, focusing on a single stage of the machine learning pipeline, either pre-processing, in-processing, or post-processing (Jain, Huber, R. Elmasri, et al. 2020; A. Biswas and Mukherjee 2021; McCormack, Bendechache, et al. 2025). This fragmented approach overlooks the interconnected ways in which bias arises and propagates across the pipeline, limiting the effectiveness of individual interventions.

2. **Fairness-accuracy trade-offs.** Many fairness-enhancing methods improve certain fairness metrics at the expense of predictive accuracy, or vice versa (M. Zhang and Sun 2022; Agarwal et al. 2018; Kinney 2025). In high-stakes domains such as criminal justice, healthcare, finance, these trade-offs pose particular problems, as prediction errors produce unjust detention or premature release. Stakeholders require clarity on how different strategies balance fairness and accuracy in practice.

3. **Data imbalance and subgroup underrepresentation.** Recidivism datasets typically suffer from both class imbalance (recidivist vs. non-recidivist outcomes) and group imbalance (underrepresentation of certain demographic groups such as women or minority populations). These imbalances skew model learn-

ing processes and contribute to systematic disparities in outcomes (Inocêncio Júnior et al. 2025; Rančić, Radovanović, and Delibašić 2021; Kabir et al. 2024; Sonoda 2023). Existing oversampling approaches partially address class imbalance but fail to account for fairness across protected groups.

4. **Hidden harms in intersectional subgroups.** Most fairness evaluations examine attributes such as race or gender in isolation. However, individuals occupy multiple marginalized identities simultaneously (e.g., Black women), where compounded disadvantages remain overlooked in single-axis analyses. This phenomenon, known as fairness gerrymandering (K. S. Andrews 2025; Kearns et al. 2018), conceals harms that disproportionately affect intersectional subgroups (Ghosh, Genuit, and Reagan 2021; Lett et al. 2025). Current fairness interventions provide limited guidance for addressing these intersectional challenges.

5. **Limited generalizability across domains.** Much of the fairness literature focuses on benchmark datasets (e.g., COMPAS (Angwin et al. 2016a), Adult Income (Chakrabarty and S. Biswas 2018)) and evaluates methods within narrow, domain-specific contexts. Researchers possess a limited understanding of how fairness-improving techniques generalize across datasets with different demographic compositions or transfer to adjacent high-stakes domains such as healthcare, where fairness concerns remain equally critical (Caton and Haas 2023; Laakom, H. Chen, Jurgen Schmidhuber, et al. 2025; Laakom, H. Chen, Jürgen Schmidhuber, et al. 2025).

These challenges highlight a critical gap between the theoretical promise of fairness-aware machine learning and its practical realization in real-world criminal justice settings. Without integrated, data-sensitive, and intersectionally aware methods, algorithmic risk assessment tools risk perpetuating inequities rather than alleviating them.

This thesis addresses these problems by developing, evaluating, and generalizing

integrated fairness pipelines that combine interventions across the machine learning lifecycle; by systematically investigating fairness-aware oversampling strategies under different imbalance conditions; and by introducing an intersectional auditing framework that exposes hidden subgroup harms. While the generalization of fairness-enhancing techniques remains a key challenge, our proposed framework was thoroughly tested on recidivism as a use case and further extended to other domains such as finance and healthcare. In doing so, this work strengthens both the technical rigor and societal trustworthiness of AI systems used for recidivism prediction and related high-stakes applications.

1.3 Research Aims and Objectives

The challenges identified in the problem statement demonstrate that existing fairness interventions in recidivism prediction remain fragmented, trade fairness for accuracy, and overlook the impacts of data imbalance and intersectionality. To address these gaps, this thesis advances the design and evaluation of fairness-aware machine learning methods that maintain technical rigor and ensure practical relevance to high-stakes domains.

The overarching aim of this thesis is to **develop and empirically validate pipeline-level approaches for building fairer, more trustworthy AI models for recidivism prediction, with particular attention to intersectional impacts and real-world data imbalance, and to derive insights that are generalizable to adjacent domains such as healthcare and finance**.

This central aim is operationalized through the following specific objectives:

1. **Conceptual and contextual grounding**: Synthesize the ethical, sociotechnical, and technical challenges of applying AI to recidivism prediction, situating fairness as a requirement for trustworthy AI in high-stakes domains.

2. **Pipeline integration**: Design and evaluate integrated fairness pipelines that combine pre-processing, in-processing, and post-processing techniques. Use

multi-objective and bi-objective optimization to characterize trade-offs between fairness and accuracy and identify Pareto-efficient configurations.

3. **Data-centric fairness**: Systematically investigate fairness-aware oversampling strategies under different conditions of class and demographic imbalance. Derive guidance on when such strategies improve fairness and accuracy and when they risk introducing synthetic bias.

4. **Intersectional auditing**: Propose and apply a practical auditing framework that evaluates fairness across intersectional subgroups (e.g., race $\times$ gender), highlighting hidden harms that are often obscured in aggregate analyses.

5. **Generalization and practice**: Assess the portability of the proposed approaches across benchmark datasets (COMPAS (Angwin et al. 2016a), RisCanvi (Jurídics i Formació Especialitzada (CEJFE) 2020), Adult Income (Chakrabarty and S. Biswas 2018)) and a private Irish Health Insurance dataset. Develop practitioner-oriented recommendations for responsible deployment of fairness-aware models.

Together, these objectives establish a roadmap for addressing the research gaps identified in the problem statement. They also provide the foundation for the specific research questions outlined in the next subsection.

1.4 Research Questions

The research aim and objectives outlined in the preceding section imply a central hypothesis that guides this work:

> **Research Hypothesis:** Integrating fairness-enhancing interventions across multiple stages of the machine learning pipeline, and complementing them with data-centric oversampling strategies and intersectional auditing, can improve fairness in recidivism prediction without

substantially compromising predictive accuracy, and these methods can be generalized to other high-stakes domains such as healthcare.

This hypothesis reflects the assumption that bias and inequity cannot be adequately mitigated by isolated, single-stage interventions, and that a holistic, pipeline-level approach provides stronger guarantees of fairness while maintaining performance.

Building on this hypothesis, the thesis is structured around a set of broad research questions (RQs), each addressing a specific dimension of fairness and trustworthiness in AI-based recidivism prediction. These overarching questions serve as conceptual anchors that link the different chapters of the thesis. While they provide the overall direction of inquiry, each chapter also contains its own sub-questions or focused investigations that delve deeper into specific technical, ethical, or methodological aspects of the broader themes.

- **RQ1**: What are the key ethical and trustworthiness challenges in AI-based recidivism prediction? See Chapter 2 and 3
  - **RQ1(a)**: What insights can be drawn from the systematic review of trustworthy AI?
  - **RQ1(b)**: What are the limitations of fairness-improving techniques in AI-based recidivism prediction across different AI development phases?
- **RQ2** : Can an integrated fairness-improving approach across different stages of AI development enhance fairness and predictive accuracy in recidivism prediction models? See Chapter 5
  - **RQ2(a)**: What are the most effective integrations of fairness-improving techniques for achieving multiple fairness metrics?
  - **RQ2(b)**: What are the practical implications of applying fairness-aware AI systems to real-world recidivism prediction?

  - **RQ2(c)**: How can the proposed integrated fairness-improving approach be generalized to other real-world recidivism datasets?
  - **RQ2(d)**: How does data imbalance affect the proposed integrated fairness-improving approach?

- **RQ3**: Can fairness-aware oversampling techniques effectively mitigate systemic biases while maintaining predictive accuracy across diverse recidivism datasets? See Chapter 6

- **RQ4**: To what extent can the proposed integrated fairness-improving approach be generalized to other high-stakes domains beyond criminal justice, such as healthcare? See Chapter 7

These research questions provide the operational framework for the experimental investigations, analyses, and contributions presented in this thesis. They also ensure that the work addresses both domain-specific concerns in criminal justice and broader challenges of fairness and trustworthiness in AI.

1.5 Thesis Contributions

This thesis makes several contributions to the field of fairness-aware machine learning, with a focus on recidivism prediction in the criminal justice system and its generalization to adjacent high-stakes domains such as healthcare. These contributions are summarized as follows:

1. **Integrated fairness pipelines**: A systematic design and empirical evaluation of multi-phase fairness pipelines that integrated pre-processing, in-processing, and post-processing interventions. The thesis provides evidence of when integrated approaches outperform single-stage interventions and maps which fairness metrics respond most strongly to different technique combinations.

2. **Optimization-informed trade-offs**: A many-objective (accuracy, Statistical Parity Difference, Disparate Impact, Equal Opportunity Difference, Predictive Equality Difference) and bi-objective optimization framework that exposes Pareto-efficient configurations. This framework makes the trade-offs between fairness and accuracy explicit, enabling stakeholders to select models that align best with context-specific priorities.

3. **Fairness-aware oversampling strategies**: A comparative study of oversampling techniques that address both class imbalance and subgroup underrepresentation. The thesis develops recommendations that specify when different oversampling strategies improve fairness, when they risk introducing synthetic bias, and how their effectiveness varies across imbalance conditions and algorithmic models.

4. **Intersectional auditing framework**: A practical framework for evaluating fairness across intersectional subgroups (e.g., race $\times$ gender). This framework identifies hidden harms that aggregate analyses often mask and demonstrates how intersectional auditing reshapes conclusions about fairness in recidivism prediction.

5. **Cross-dataset generalization**: A set of cross-dataset experiments (COMPAS, RisCanvi, Adult Income, Health Insurance) that evaluate the robustness of fairness-enhancing techniques across datasets with different demographic compositions. These analyses provide insights into the limits of generalizability in fairness-aware machine learning.

6. **Domain transfer to healthcare**: An extension of the proposed methods to a real-world Irish healthcare dataset, illustrating how the fairness pipelines, oversampling strategies, and intersectional auditing framework adapt to real-world settings. This contribution demonstrates both the portability and the contextual limitations of the approaches beyond criminal justice.

7. **Practitioner guidance**: A set of actionable recommendations for researchers,

policymakers, and practitioners on the responsible deployment of fairness-aware models. These guidelines address real-world challenges such as class imbalance, small subgroup sizes, incompatible fairness metrics, and the risk of overcompensation or synthetic bias in data-centric approaches.

Collectively, these contributions advance the state of the art by moving from isolated, single-stage fairness interventions toward integrated, pipeline-level strategies. They also provide practical tools and recommendations for improving fairness in domains where algorithmic decisions carry significant ethical and societal consequences.

1.6 Thesis Structure

The remainder of this thesis organizes into seven chapters. Each chapter builds upon the preceding one to develop a coherent argument and set of contributions around fairness-aware machine learning in recidivism prediction and its generalization to other high-stakes domains.

- **Chapter 2: Background: Foundations of Trustworthy AI and Fairness-Aware Artificial Intelligence** This chapter provides the conceptual and technical foundations necessary for understanding trust, trustworthy AI and fairness-aware AI. It reviews the different principles of trustworthy AI in high-stakes domains, key concepts of bias and fairness, definitions of fairness metrics, the role of intersectionality, and the structure of machine learning pipelines. It concludes by situating fairness within broader discussions of trust in AI.

- **Chapter 3: Literature Review: Trustworthy AI, Algorithmic Fairness, and Intersectionality** This chapter critically examines prior research on trustworthy AI, AI fairness and Intersectionality with a focus on recidivism prediction. It synthesizes fairness-improving interventions across the

ML pipeline and identifies key limitations, such as overreliance on isolated techniques and lack of intersectional evaluation. This review motivates the integrative approaches proposed in the thesis.

- **Chapter 4: Methodology: Multi-Phase Fairness Pipelines and Optimization Techniques** This chapter details the research design and experimental setup: datasets (COMPAS, RisCanvi, Adult Income, and the Irish healthcare dataset), task definitions, sensitive attributes and subgroup construction, performance and fairness metrics (e.g., SPD, DI, EOD, PED), model families, integrated pipeline configurations (pre-, in-, post-processing), optimization procedures (many-objective and bi-objective), validation protocols, statistical testing, and ethical considerations (governance, transparency, and reproducibility).

- **Chapter 5: Fairness Pipelines Integration: Technical and Empirical Investigations** This chapter develops and evaluates integrated fairness pipelines that integrates pre-processing, in-processing, and post-processing methods. Using multi-objective optimization, it identifies Pareto-efficient models that balance fairness and accuracy. Results across the COMPAS and RisCanvi datasets illustrate the strengths and limitations of integrated approaches compared to single-stage methods.

- **Chapter 6: Effect of Data Oversampling Techniques on Fairness and Performance** This chapter investigates fairness-aware oversampling strategies for handling class and subgroup imbalance. It systematically compares traditional and fairness-aware oversampling techniques across multiple classifiers and fairness metrics.

- **Chapter 7: Intersectional Fairness: Auditing Compound Bias in Algorithmic Decisions** This chapter extends the proposed approaches beyond criminal justice. It evaluates the generalizability of fairness pipelines and oversampling strategies on the Adult Income dataset and applies the intersec-

tional auditing framework to a real-world Irish healthcare Insurance dataset. The findings highlight both the potential and the limitations of transferring fairness-aware methods to other high-stakes domains.

- **Chapter 8: Discussion and Conclusion: Contributions, Implications and Future Directions** The final chapter summarizes the main contributions of the thesis, reflects on its limitations, and outlines directions for future research. It emphasizes the importance of integrating technical, ethical, and sociotechnical considerations in building trustworthy AI for high-stakes decision-making.

This structure ensures a logical progression from conceptual foundations and systematic literature review, through methodological innovation, to empirical validation and cross-domain generalization. It provides a clear roadmap for how the thesis addresses its research questions and delivers its contributions.

Chapter 2

Background: Foundations of Trustworthy AI and Fairness-Aware Artificial Intelligence

2.1 Chapter Overview

Artificial Intelligence (AI) systems increasingly shape high-stakes decision-making in domains such as criminal justice, healthcare, and finance, where their societal implications remain profound (Montani and Striani 2019). While these technologies promise scalability and consistency, they also raise significant ethical concerns, including accountability, fairness, robustness, explainability, transparency, privacy, and human oversight. Among these, fairness attracts particular attention from stakeholders in high-stakes domains. A central challenge lies in determining whether such models ensure equity and non-discrimination across individuals and demographic groups, given that structural biases and inequalities often permeate the data and models themselves.

This chapter provides the conceptual and technical foundations for understand-

ing fairness-aware machine learning within the broader context of Trustworthy AI. Section 2.2 explores how AI systems influence decision-making in sensitive domains, while Section 2.3 examines the concept of trust, the different types of trust relevant to AI, and the principles underpinning trustworthy AI. Section 2.4 analyzes the sources and forms of algorithmic bias. Section 2.5 discusses fairness as a core element of trustworthy AI and outlines fairness interventions across the machine learning pipeline. At the same time, Section 2.6 explores the various fairness metrics and the potential trade-offs among them. Section 2.7 introduces the concept of intersectionality as a theoretical lens for analyzing overlapping identities and compounding disadvantage (Crenshaw 2022). Finally, Section 2.8 concludes the chapter by reiterating the conceptual and definitional foundations that inform the rest of this thesis.

Collectively, these sections provide the theoretical scaffolding for this thesis. They ground fairness in both formal statistical definitions and socio-technical contexts, while highlighting the challenges of metric incompatibility, fairness–accuracy trade-offs, and generalizability. These challenges motivate the integrated methodological framework developed in Chapter 4.

2.2 Algorithmic Decision-Making in High-Stakes Domains

The increasing deployment of AI systems in high-stakes decision-making domains raises critical questions regarding fairness, justice, and trustworthiness. These domains, such as criminal justice, healthcare, and finance, faces decisions that significantly affect individuals' lives, liberties, and access to opportunities (Montani and Striani 2019). In such contexts, algorithmic predictions do not remain neutral; they reflect and often reinforce historical patterns of inequality embedded in the data, institutional practices, and broader socio-political systems (Eubanks 2018; Buolamwini and Gebru 2018; Angwin et al. 2019). This pattern is evident in the

deployment of widely used recidivism risk assessment tools, including COMPAS and RisCanvi, which illustrate both the potential and the challenges of algorithmic decision-making in criminal justice.

In the healthcare sector, similar patterns of harm emerge. Risk prediction algorithms used for patient triage and treatment allocation underestimate the needs of Black patients when proxies such as historical healthcare expenditure serve as input features (Obermeyer et al. 2019). Such proxies implicitly encode structural inequities in access to care (Obermeyer et al. 2019). In education, dropout prediction models penalize students from under-resourced schools, effectively perpetuating existing educational disparities (Keyes, Hutson, and Durbin 2019). In finance and hiring, algorithms trained on biased historical data reproduce discriminatory lending practices or employment screening outcomes (Fabris et al. 2025).

These examples illustrate that algorithmic decision-making in high-stakes domains cannot exist in isolation from the social and historical conditions in which they operate. As recent literature emphasizes, algorithmic systems represent not merely technical artifacts but sociotechnical interventions (Green and Y. Chen 2019; Birhane 2021). When institutions deploy such systems without careful consideration of context, they risk exacerbating the very inequalities they aim to resolve.

Another critical concern arises from the opacity of these systems. Many risk assessment tools, especially those developed by the private sector, operate as black-box models with proprietary algorithms and limited transparency (Lo Piano 2020; McKay 2020). This opacity raises critical concerns about accountability, transparency, and explainability. If an algorithm contributes to a wrongful decision—such as the unjust denial of parole—who bears responsibility: the developer, the agency deploying the tool, or the criminal justice officials relying on its output? Moreover, how can the decisions generated by these AI systems be explained or interpreted in ways that are understandable and defensible? These unresolved questions reveal persistent governance gaps and underscore the urgent need for robust ethical frameworks and regulatory oversight to ensure accountability in the deployment of algo-

rithmic systems (Figueroa-Armijos, B. B. Clark, and Motta Veiga 2022; Mökander et al. 2022).

In response to these concerns, frameworks for "Trustworthy AI" have emerged, particularly in the European context. The European Commission's Ethics Guidelines for Trustworthy AI outline key requirements such as technical robustness and safety, privacy and data governance, diversity, non-discrimination and fairness, transparency, accountability, societal and environmental well-being, and human agency and oversight (H. AI 2019; European Parliament and Council 2024). However, research shows that existing AI tools often fall short of meeting these requirements, particularly in their treatment of marginalized groups (R. Berk, Heidari, Jabbari, Kearns, et al. 2021; Ződi 2022). Studies also reveal a lack of consensus on what constitutes fairness in these systems, with competing definitions leading to trade-offs that policy and practice rarely acknowledge (R. Berk, Heidari, Jabbari, Kearns, et al. 2021; Grgic-Hlaca et al. 2018).

In summary, algorithmic decision-making in high-stakes domains requires a multidimensional understanding of the ethical imperatives underpinning trustworthy AI. Addressing these imperatives demands both technical and non-technical solutions that account for broader societal and contextual challenges. The following section explores the anatomy of trustworthy AI and identifies key entry points through which ethical principles integrate throughout the AI system lifecycle.

2.3 Trust and Trustworthy AI

This section provides an overview of the concept of trust, the different types of trust relevant to AI, and the principles underlying trustworthy AI.

2.3.1 Trust

Trust serves as a foundational component for the successful adoption and acceptance of AI systems, especially in high-stakes domains (Alikhademi et al. 2021;

O'Loughlin and Bukowitz 2021; Connolly 2013). However, scholars regard trust in AI as a complex, multifaceted concept that draws significant attention across disciplines, including computer science, philosophy, law, and sociology (Ryan 2020). Researchers explore trust's antecedents, types, interpersonal dynamics, and its ethical and practical implications, yet they have not established a universally accepted definition.

In the context of human–AI interaction, scholars define trust as a psychological or behavioral state that involves the willingness of a trustor (e.g., a user or stakeholder) to rely on a trustee (e.g., an AI system) in situations that involve uncertainty or risk (Connolly 2013; Beshi and R. Kaur 2020). Classical theories distinguish between various types of trust. Scholars identify several typologies of trust that describe how and why individuals place confidence in others, including artificial systems. The following types of trust hold particular relevance in the context of human–AI interaction (Connolly 2013):

- **Attitudinal trust**: This form of trust rests on an individual's internal beliefs, values, and worldviews. It reflects a general disposition to trust others (including institutions or technologies) based on prior experience, cultural background, or psychological inclination. In the context of AI, attitudinal trust manifests when users adopt a positive or skeptical stance toward AI systems based on their perceptions of science, authority, or technology, rather than direct interaction with the system itself. For example, individuals with high institutional trust tend to trust a government-deployed algorithm, even in the absence of technical transparency.

- **Predictability-based trust**: This type of trust emerges from consistent and reliable behavior over time. Individuals build it through repeated interactions in which the AI system demonstrates stable, expected performance. Predictability-based trust holds particular relevance in AI applications that affect individuals repeatedly or longitudinally, such as predictive policing, credit scoring, or medical diagnostics. When systems behave consistently and

make decisions that align with stakeholder expectations, individuals accumulate trust. However, erratic or opaque behavior, even when accurate, erodes this form of trust.

- **Voluntarist trust**: Voluntarist trust involves a conscious decision by the trustor to accept vulnerability in anticipation that the trustee acts benevolently or fairly. It reflects a moral or psychological commitment to trust, despite potential risks or asymmetries of power. In AI systems, individuals demonstrate voluntarist trust when they subject themselves to algorithmic decision-making (e.g., automated hiring systems or risk assessments) without full visibility into how those systems operate, assuming or hoping that the system remains fair, ethical, and aligned with their interests. This kind of trust depends on broader institutional guarantees such as regulatory oversight or ethical certification.

These views converge on the idea that trust entails a level of risk and vulnerability. For example, in recidivism prediction tools, individuals subject to algorithmic assessments rely on the system's fairness, accuracy, and transparency, often without control over its operations (Thiebes, Lins, and Sunyaev 2021). In this dynamic, the AI system acts as the trustee, while citizens and institutions serve as trustors, earning the system's confidence.

Several key qualities influence whether stakeholders choose to trust an AI system. These characteristics, often referred to as the *antecedents of trust*, help determine the perceived trustworthiness of the system from the perspective of users and affected individuals (K. Liu and Tao 2022; Toreini et al. 2020).

- **Ability**: This concept refers to the AI system's perceived competence and technical capacity to carry out its intended tasks effectively. In the context of AI, this competence includes accurate predictions, reliable decision-making, and alignment with performance expectations. A system that consistently meets or exceeds benchmarks in its domain, such as diagnosing disease accu-

rately in medical AI or assessing risk reliably in criminal justice, appears more capable and, thus, more trustworthy.

- **Benevolence**: Benevolence captures the extent to which the AI system (or, more precisely, its designers and operators) acts in the interest of users or society rather than for self-serving purposes. Although AI lacks intention per se, stakeholders assess whether the motivations behind the system's development and deployment align with the public good. For instance, an AI used in public policy that incorporates fairness constraints or participatory design processes appears more benevolent.

- **Integrity**: Integrity refers to the perception that the AI system adheres to a coherent set of ethical principles, norms, or values over time. It implies a commitment to honesty, fairness, and moral consistency, both in system behavior and in the governance structures that oversee its use. For example, an AI system that maintains consistent criteria or respects privacy regulations demonstrates integrity.

- **Predictability**: Predictability involves the consistency and stability of the AI system's outputs when it processes similar or equivalent inputs. A system that produces volatile or contradictory decisions, such as granting a loan to one applicant but denying another with the same profile, undermines user confidence. Predictability reassures stakeholders that the system behaves in an expected and understandable manner across time and situations.

Moreover, trust operates as a context-sensitive construct mediated by human, environmental, and technological factors (Toreini et al. 2020). Human factors include cultural background, values, and previous experiences with institutions or technologies. Environmental influences encompass governance structures, education systems, and public discourse, while technological factors concern system performance, transparency, and reliability.

2.3.2 What Is Trustworthy AI?

Because trust in AI must be earned rather than assumed, the concept of *trustworthy AI* emerges to guide the ethical development and deployment of intelligent systems. Trustworthiness refers to the ability of a system to justify the confidence stakeholders place in it; it reflects a combination of competence, integrity, and alignment with human values (Sutrop 2019).

The European Commission's High-Level Expert Group on AI (HLEG AI) lays a foundational framework in its 2019 report (H. AI 2019), which defines trustworthy AI as:

> *"AI that is lawful, ethical, and robust, with these three components being mutually reinforcing."*

The HLEG outlines four ethical principles: respect for human autonomy, prevention of harm, fairness, and explicability. It also identifies seven concrete requirements: Human agency and oversight, Technical robustness and safety, Privacy and data governance, Transparency, Diversity, non-discrimination, and fairness, Societal and environmental well-being, and Accountability.

These requirements serve as a baseline for both public and private sector actors, though they remain non-exhaustive. Hence, this thesis extends these principles better to accommodate domain-specific contexts, such as recidivism prediction, as illustrated in Figure 3.2.

2.3.3 Trust, Risk, and the AI Lifecycle

The relationship between trust and trustworthiness in AI systems functions as a reciprocal dynamic: users must choose to take a risk (i.e., trust), while developers and institutions must ensure that the system remains genuinely worthy of that trust (i.e., trustworthy). This distinction holds critical importance, especially in domains where users experience decisions without meaningful input or recourse.

Designers and regulators must cultivate trust through deliberate design and over-

sight mechanisms embedded throughout the AI lifecycle, from data collection and model training to deployment, monitoring, and redress. Because trust operates as a dynamic and context-dependent construct, practitioners must conduct ongoing evaluation, engage stakeholders, and apply participatory design to maintain public trust in AI technologies.

In conclusion, building and sustaining trust in AI systems requires aligning technological capabilities with ethical principles, legal safeguards, and socio-institutional expectations. Only when these dimensions converge can developers and institutions achieve truly trustworthy AI.

2.3.4 Principles of Trustworthy AI

Trustworthy AI refers to systems that remain lawful, ethical, and technically robust, as frameworks from the European Commission (H. AI 2019), the OECD (Yeung 2020), and academic research (Jobin, Ienca, and Vayena 2019) articulate. To earn public trust and ensure responsible deployment, AI systems must adhere to a multidimensional set of principles that span ethical, technical, and societal domains. The sections below explore these principles individually, fairness, accountability, transparency, robustness, privacy, human oversight, and societal well-being, each of which contributes uniquely to the legitimacy and acceptance of AI in high-stakes environments, especially within the criminal justice system when predicting recidivism risk.

- **Accountability:**

  Accountability in AI refers to the principle that those who design, develop, deploy, oversee, or use AI systems must remain answerable for the consequences of those systems' decisions and behaviors. This principle ensures that AI does not operate in a vacuum but remains embedded within institutional, legal, and ethical frameworks that enable responsibility, redress, and governance (Jobin, Ienca, and Vayena 2019; Raji and Buolamwini 2019).

Unlike conventional software, AI systems evolve, operate in an opaque manner, and produce unintended outcomes. Therefore, stakeholders must understand accountability as both a technical and institutional commitment. It involves assigning responsibility for the design and development of the system (e.g., choices in data, features, training objectives), the deployment context (e.g., who uses the model, under what conditions), and the post-deployment impact (e.g., errors, biases, or harms affecting end users or communities). Key mechanisms that support accountability include:

- **Auditing and Evaluation:** Organizations conduct regular audits, both internal and third-party, to assess whether models perform fairly and reliably across demographic groups. These audits include bias audits, explainability assessments, and security tests. Raji et al. (2020) emphasize that actionable auditing practices play a crucial role in detecting harms before they scale (Raji, Smart, et al. 2020).
- **Algorithmic Impact Assessments (AIAs):** Drawing inspiration from environmental and privacy impact assessments, AIAs serve as structured procedures that evaluate the ethical, legal, and social implications of an AI system before and during deployment. They compel organizations to anticipate risks, engage stakeholders, and justify design choices (Reisman et al. 2018).
- **Redress and Contestability Mechanisms:** A key dimension of accountability ensures that affected individuals can challenge decisions and seek recourse. This principle aligns with democratic norms and legal obligations such as the GDPR's "right to explanation" (Voigt and Von dem Bussche 2017). Contestability requires accessible appeal procedures, human-in-the-loop review systems, and public complaint channels.
- **Role Assignability and Traceability:** Accountability depends on traceable responsibility, meaning that organizations must identify who

participates in each stage of the pipeline. This practice includes logging model version histories, documenting data lineage, and defining role-based access controls.

- **Public Transparency and Reporting:** Public institutions and companies that deploy high-impact AI should provide clear documentation, such as model cards (Mitchell et al. 2019), datasheets for datasets (Gebru et al. 2021), or system cards, that disclose performance metrics, known limitations, and intended use cases.

In high-stakes domains such as healthcare, criminal justice, employment, and social welfare, the absence of accountability creates severe consequences, from wrongful arrests to algorithmic discrimination and financial exclusion. In such settings, designers and institutions must treat accountability not as an afterthought but as a central design and governance priority. Accountability ensures not only functional correctness but also moral legitimacy, public trust, and institutional responsibility.

Finally, emerging governance frameworks such as the EU AI Act (2024) and OECD AI Principles (2019) recognize accountability as a pillar of trustworthy AI and urge organizations to embed oversight at every stage of the AI lifecycle, from conception to deployment, monitoring, and decommissioning (Yeung 2020; European Parliament and Council 2024).

- **Fairness and Non-Discrimination:**

Fairness serves as a cornerstone of trustworthy AI and a normative imperative in systems that affect access to rights, resources, and opportunities. In the context of AI, fairness refers to the mitigation of algorithmic bias and the equitable treatment of individuals and groups, particularly those historically subjected to discrimination and marginalization. Non-discrimination requires that AI systems must not produce or exacerbate disparities based on protected characteristics such as race, gender, age, disability, sexual orientation,

or socioeconomic status (Mehrabi et al. 2021; Gohar and L. Cheng 2023).

Achieving fairness in AI represents not a singular technical goal but a multi-dimensional, context-sensitive challenge. Bias can enter the system at various stages of the machine learning pipeline (Mehrabi et al. 2021), from the choice of features and data sampling strategies to model training and post-deployment usage. Sources of unfairness often include systemic factors that reflect and reinforce structural and societal inequalities. Understanding the different types of bias is essential for identifying points of intervention across the machine learning pipeline. Key categories of bias include:

- **Historical bias:** Historical bias reflects pre-existing inequities already present in the world and appears in training datasets. Models do not introduce this bias themselves; instead, they inherit it from historically discriminatory practices, policies, or social systems. For instance, crime data reflects over-policing in minority communities, thereby encoding systemic racism into predictive policing models (Barocas, Hardt, and Narayanan 2023). Historical bias proves particularly insidious because it appears "neutral" or "factual" simply because it remains embedded in large-scale datasets.
- **Representation bias:** Representation bias occurs when certain groups remain underrepresented, misrepresented, or absent in the training data. This imbalance causes models to perform poorly or unfairly for those groups. For example, facial recognition systems trained predominantly on lighter-skinned individuals show significantly higher error rates for darker-skinned and female faces (Buolamwini and Gebru 2018). Such bias undermines model generalizability and contributes to algorithmic exclusion or harm for underrepresented populations.
- **Measurement bias:** Measurement bias arises when variables used to train or evaluate models remain inaccurate, imprecise, or serve as flawed

proxies for the outcomes of interest. In healthcare, for example, using healthcare expenditure as a proxy for health need systematically underestimates the needs of low-income and marginalized groups who spend less on care, not because they are healthier, but because they face barriers to access (Obermeyer et al. 2019). Measurement bias distorts model learning and exacerbates inequities even when prediction accuracy appears high.

- **Model bias:** Model bias occurs when the structure, learning process, or chosen hyperparameters of an algorithm introduce unfairness, even when the dataset remains unbiased or balanced (Mehrabi et al. 2021; Pessach and Shmueli 2022). This form of bias resides within the modeling pipeline and often stems from how algorithms prioritize certain features, optimize for performance, or generalize patterns during training. One common trigger of model bias arises when an algorithm assigns disproportionate weight to specific input features without justifiable reasons related to fairness or domain knowledge (D. Kaur et al. 2022; Standardization 2021; Varona and Suárez 2022).
- **Deployment bias:** Deployment bias emerges when an AI system operates in real-world contexts that deviate from the conditions assumed during design, training, or testing (Holstein et al. 2019). These deviations include shifts in population demographics, institutional practices, or environmental variables. For instance, a recidivism risk model trained on data from one jurisdiction performs poorly when deployed in another with different legal practices or social conditions. Deployment bias underscores the importance of continual model validation and contextual sensitivity.

To mitigate these biases, developers and institutions incorporate fairness-aware design into AI systems through both technical interventions and participatory practices (Delgado et al. 2023).Numerous fairness-enhancing techniques have

been proposed and fairness metrics, such as statistical parity, equal opportunity, and disparate impact, to mitigate, detect, and quantify inequality in model outcomes (Chouldechova 2017; Kleinberg, Mullainathan, and Raghavan 2016). Yet these metrics often conflict, requiring domain-sensitive trade-offs (Z. Chen et al. 2023). As a result, practitioners must implement more robust and fairness-enhancing methods holistically across preprocessing, in-processing, and post-processing phases rather than apply them in isolation (Caton and Haas 2023; Z. Chen et al. 2023).

Fairness also requires an intersectional lens (Kearns et al. 2018). Most evaluations focus on single attributes (e.g., race or gender), yet harms compound at intersections, such as for Black women, disabled queer youth, or low-income immigrant communities (Kearns et al. 2018; Birhane 2021). Intersectional frameworks and subgroup auditing aim to uncover these "hidden" disparities (A. Wang, Ramaswamy, and Russakovsky 2022; Islam et al. 2023).

Procedural and participatory dimensions remain equally important. Scholars such as Costanza-Chock 2020 and Delgado et al. 2023 argued that fairness cannot be imposed from above but must be co-constructed with affected communities through inclusive design, participatory audits, and governance mechanisms (Costanza-Chock 2020; Binns 2018).

Ultimately, fairness entails more than avoiding discrimination; it demands the active promotion of equity. This perspective requires correcting structural disadvantages and prioritizing protections for the most vulnerable. In high-stakes domains such as criminal justice, healthcare, and finance, fairness failures risk entrenching systemic injustice. Accordingly, AI systems must aim not only to "do no harm" but also to advance justice, inclusion, and human dignity (Veale, Van Kleek, and Binns 2018; Eubanks 2018).

- **Reliability:**

Reliability refers to an AI system's ability to perform its intended functions ac-

curately, stably, and dependably across diverse scenarios, data distributions, and operational conditions. Unlike consistency, which emphasizes identical outputs under identical inputs, reliability stresses robust and generalizable performance under evolving conditions such as data drift, input noise, or context shifts (Amodei et al. 2016; Floridi 2021; LEARNING 2009).

A reliable system maintains predictive accuracy and stability across populations, deployment settings, and temporal variations. For example, a recidivism prediction tool should perform consistently across jurisdictions and demographic groups and remain effective as criminal justice practices evolve. In contrast, unreliable systems often fail in underrepresented scenarios, yielding erratic or misleading outputs.

This requirement remains especially critical in high-stakes domains such as healthcare and criminal justice, where even minor reliability failures, caused by adversarial examples, rare edge cases, or distribution shifts, can produce disproportionate harm (Amodei et al. 2016; LEARNING 2009). Overfitted models, for instance, may appear reliable in training but collapse under real-world conditions, underscoring the need for rigorous validation, domain adaptation, and uncertainty estimation (LEARNING 2009).

Ultimately, reliability extends beyond average-case performance to encompass resilience under adverse real-world circumstances. A reliable AI system demonstrates stability not only in controlled environments but also when it faces the variability and unpredictability of deployment contexts. Key dimensions of AI reliability include:

- **Temporal stability or shifts**: This dimension refers to the system's ability to maintain consistent performance over time, even as the underlying data distribution evolves due to seasonal shifts, user behavior changes, or socio-political dynamics. Many models deployed in practice often suffer temporal quality degradation ("AI aging") even when input

distributions remain relatively stable (Yao et al. 2022; Murch, Kairouz, and French 2024; Cai, Namkoong, and Yadlowsky 2023; Vela et al. 2022). For example, a loan approval model should not experience sudden degradation as economic conditions fluctuate.

- **Contextual robustness**: Robustness across contexts ensures that the AI system adapts to variations in input environments, tasks, or user characteristics without severe performance loss. This includes resilience to domain shifts, adversarial conditions, and novel use cases. Hence, context robustness ensures that models develop on datasets whose feature distributions or label conditional mechanisms differ from those used in training does not drop in performance (Cai, Namkoong, and Yadlowsky 2023; Yao et al. 2022).
- **Population generalizability**: A reliable system must generalize equitably across diverse demographic groups, including those historically underrepresented. Empirical work on calibration across subpopulations demonstrates that models may display acceptable aggregate accuracy yet perform poorly for specific groups (age, race, gender, etc.) unless developers apply fairness-aware calibration or reweighting techniques (Pleiss et al. 2017; Barda et al. 2021).
- **Graceful degradation**: When the AI system encounters degraded conditions such as corrupted inputs, noisy data, sensor failures, or adversarial interference, it should fail predictably. Studies on AI aging and temporal degradation show that performance decay can occur gradually and predictably, suggesting that developers can design systems to respond gracefully rather than fail catastrophically (Vela et al. 2022; Cai, Namkoong, and Yadlowsky 2023).

Reliability serves as a precondition for trust, especially in sociotechnical systems where humans interact with or depend on AI across diverse jurisdictions. Unreliable models erode user confidence, complicate human oversight,

and raise ethical concerns about unjust or unpredictable decision-making. For these reasons, international frameworks such as the European Commission's Ethics Guidelines for Trustworthy AI (H. AI 2019; Floridi 2021) and the OECD AI Principles (OECD 2021; Yeung 2020) emphasize reliability as a foundational principle for ethical AI development and deployment.

Ensuring reliability requires a variety of technical strategies, including regular model retraining with fresh data, stress-testing under edge cases, domain generalization methods, calibration, and ensemble approaches. Institutions must complement these strategies with safeguards such as post-deployment monitoring, user feedback loops, and contingency plans for system failures to uphold reliability in dynamic real-world settings.

- **Technical Robustness and Safety:**

  Technical robustness and safety refer to an AI system's ability to function reliably and predictably under a wide array of conditions, including stress, attack, uncertainty, adversarial manipulation, and operational changes. This principle encompasses a system's resilience to failure, its ability to handle noise or perturbations, and its preparedness for edge-case scenarios that developers did not include explicitly in the training data (Amodei et al. 2016; Mitchell et al. 2019).

  Robustness focuses on the system's resistance to internal and external perturbations or security attacks. For instance, a robust AI model for credit scoring or criminal risk prediction should produce stable outputs even when it encounters noisy inputs, minor feature manipulations, or distributional shifts between training and deployment environments. It must also withstand attempts at manipulation via adversarial attacks, where small, carefully crafted changes to the input lead to disproportionately erroneous outcomes (Szegedy et al. 2013).

  Safety, by contrast, relates to minimizing harm and ensuring that the system behaves within acceptable bounds, especially in high-stakes or safety-

critical applications such as autonomous driving, healthcare, or law enforcement. Safe AI systems anticipate potential hazards, fail gracefully when limits are exceeded, and incorporate redundancy or human-in-the-loop mechanisms to prevent catastrophic consequences (Amodei et al. 2016). For example, an autonomous vehicle must not only avoid collisions during normal operation but also respond safely during hardware failure or uncertain perception scenarios.

Technical robustness and safety together capture the capacity of AI systems to maintain stable, secure, and reliable behavior under a range of challenging conditions, including data shifts, adversarial attacks, or unexpected operational scenarios. These properties remain essential for preventing harm and ensuring trustworthy deployment, particularly in high-stakes environments. Key aspects include:

- **Adversarial robustness**: This concept refers to the AI system's resilience against adversarial inputs, carefully crafted, often imperceptible perturbations to input data that cause the model to make incorrect or harmful predictions (Goodfellow, Shlens, and Szegedy 2014). For instance, a subtle modification to an image can mislead a classifier into misidentifying a stop sign as a speed limit sign in autonomous driving systems. In the criminal justice context, a minor alteration to a defendant's record (e.g., shifting the number of prior convictions from two to one) could lower a recidivism risk score without reflecting any real change in underlying behavior. Developers build adversarial robustness through techniques such as adversarial training, input sanitization, and uncertainty estimation.
- **Data drift detection**: Over time, the statistical properties of incoming data may shift away from the distribution the model was trained on, a phenomenon known as concept or data drift. Drift can arise from seasonal trends, policy changes, or shifts in user behavior. Robust AI systems must incorporate mechanisms for detecting such drift and either adapt

the model or trigger human review. Failure to detect drift can result in silent model decay and increase error rates over time. Works such as Bayram, Ahmed, and Kassler 2022; Choudhary et al. 2022 and Pham et al. 2025 provide taxonomies and detection methods for managing this phenomenon.

- **Model degradation handling**: As models age, their performance may deteriorate due to evolving real-world contexts or outdated training assumptions. Technical robustness requires systems to monitor their own performance and handle degradation gracefully. This includes retraining protocols, confidence-based abstention mechanisms, and meta-learning strategies that enable models to self-assess and maintain acceptable service quality over time. The survey by Bayram, Ahmed, and Kassler 2022 and studies like Choudhary et al. 2022 illustrate how developers can anticipate and manage degradation effectively.
- **Fail-safe design**: A fail-safe system includes engineered safeguards that activate when the model encounters uncertain, contradictory, or hazardous input conditions. These safeguards may involve fallback procedures such as switching to a rule-based backup system, alerting a human operator, or suspending the automated decision process. Such mechanisms remain critical in applications like healthcare diagnostics or autonomous vehicles, where unchecked system failure can result in significant harm. The work by G. A. Lewis et al. 2022 includes production monitoring and alerts as part of robustness infrastructure.
- **Out-of-Distribution (OOD) generalization**: In real-world deployment, AI systems frequently encounter data points that differ significantly from the training distribution. These points may include new classes, unforeseen edge cases, or anomalous behaviors. A robust system must recognize when an input is out of distribution and respond appropriately, either by flagging it for human review, abstaining from prediction, or

using uncertainty-aware models. Failure to address OOD scenarios can lead to unpredictable or biased outcomes. The surveys by J. Liu et al. 2021 and Jingkang Yang et al. 2024 provide comprehensive frameworks and empirical findings for addressing these challenges.

Robustness and safety stand at the center of public trust in AI systems. A single failure in a sensitive application, such as misdiagnosing a medical condition, mislabeling a low-risk individual as high-risk, or failing to detect a pedestrian, can cause irreversible harm and provoke public backlash. Hence, ensuring robustness serves as not only a technical imperative but also a moral and social one.

Developers and institutions deploy various technical strategies to support these goals, including adversarial training, ensemble models, uncertainty quantification, robust loss functions, and continuous post-deployment monitoring. They must embed these strategies throughout the AI development pipeline, from data preprocessing to deployment and system retirement.

Moreover, AI guidelines and frameworks explicitly require technical robustness and safety as core criteria for the development and deployment of trustworthy AI. These frameworks stress the importance of proactive risk assessment, ethical foresight, and contingency planning.

Ultimately, robust and safe AI systems not only maximize performance under ideal conditions but also minimize harm under imperfect or adversarial scenarios. As AI operates increasingly in real-world, dynamic environments, these attributes become indispensable for long-term sustainability, legal compliance, and ethical acceptability.

- **Explainability, Transparency, and Interpretability:** These three interrelated yet conceptually distinct principles serve as central pillars for building trustworthy AI systems. Together, they make machine learning (ML) models and artificial intelligence (AI) systems more understandable, auditable, and

aligned with human values, especially in high-stakes decision-making.

- **Transparency** refers to the openness and accessibility of information regarding the structure, functioning, and data inputs of an AI system. A transparent AI system provides visibility into its model architecture (e.g., whether it uses decision trees or deep neural networks), data provenance (e.g., sources and quality of training data), performance metrics, hyperparameter settings, and known limitations (Mitchell et al. 2019; European Parliament and Council 2024). Transparency also extends to organizational and procedural elements, including who builds the model, its purpose, and the ethical guidelines under which it is developed. It lays the foundation for scrutiny, compliance audits, and public trust.
- **Explainability** concerns the ability of an AI system to generate comprehensible justifications for its outputs. It seeks to answer questions such as "Why did the system make this prediction or decision?" Explainability plays a crucial role in contexts such as healthcare, criminal justice, and finance, where individuals affected by algorithmic decisions have a right to understand the rationale behind them (Doshi-Velez and B. Kim 2017). Explainability techniques include saliency maps, counterfactual explanations, and feature attribution methods (e.g., SHAP, LIME) (Delaney et al. 2023; Ghnemat, Alodibat, and Abu Al-Haija 2023). Importantly, Explainability focuses on providing post hoc, human-understandable reasoning, even when the underlying model remains a black box.
- **Interpretability**, in contrast, relates to how inherently understandable the internal mechanisms of a model are. It focuses on the degree to which a human can comprehend the actual mathematical or logical structure of the model without additional tools. Interpretable models, such as linear regression, decision trees, or rule-based systems, offer transparent decision boundaries by design. According to Lipton (Lipton 2018), Interpretability includes both simulatability (can a human mentally simulate

the model?) and decomposability (can each part of the model be understood intuitively?).

Together, explainability, transparency, and interpretability enhance stakeholder trust and allow for meaningful oversight. They support key AI governance practices such as:

- **Debugging and error analysis**: By enabling practitioners to trace how decisions are made, transparent and interpretable systems allow them to identify flaws, biases, or data artifacts more effectively.
- **Fairness auditing and bias detection**: Transparent and interpretable systems facilitate fairness evaluations across demographic groups, making bias detection and mitigation more systematic.
- **Regulatory compliance**: Regulations such as the EU AI Act and GDPR emphasize the "right to explanation" in automated decision-making. Systems that lack transparency or explainability risk violating these legal requirements.
- **User trust and social legitimacy**: Users tend to accept and adopt AI systems whose behaviors they can understand, question, and contest, thereby enhancing public trust and legitimacy.

Despite their interconnected nature, researchers and practitioners must recognize that high performance in one dimension does not guarantee strength in the others. For example, a deep neural network may appear partially explainable when practitioners use tools such as LIME, but it remains fundamentally uninterpretable and opaque in its internal workings. Hence, these principles must be pursued in tandem, tailored to the application's context, and integrated into system design from the outset rather than added retroactively.

- **Societal and Environmental Wellbeing:**

Trustworthy AI must actively contribute to human flourishing while safeguarding the sustainability of the environments in which it operates. The principle of societal and environmental well-being emphasizes that AI systems should promote the common good, minimize harm, and enhance collective outcomes across social, economic, cultural, and ecological domains (Jobin, Ienca, and Vayena 2019; Floridi 2021).

From a societal perspective, AI should reinforce democratic values, social cohesion, and inclusion. This entails preventing technologies that deepen inequality, marginalize vulnerable groups, or erode social trust. In high-impact areas such as education, healthcare, housing, and public welfare, AI must expand equitable access rather than reproduce structural disadvantages, for example, mitigating bias in hiring systems or supporting personalized learning without worsening digital divides.

AI deployment must also consider its broader systemic effects. Algorithmic decision-making in transportation, finance, or criminal justice requires continuous monitoring for unintended consequences such as job displacement, algorithmic redlining, or mass surveillance, since these risks directly affect public trust and institutional legitimacy (Yeung 2020).

On the environmental side, the growing scale of deep learning introduces significant carbon costs. Researchers such as Strubell et al. (Strubell, Ganesh, and McCallum 2020) found that training a single large language model can emit as much carbon as five cars do over their entire lifetimes. Sustainable AI development, therefore, demands energy-efficient algorithms, model compression, low-power hardware, and carbon-aware data centers. Governments, industry, and researchers increasingly acknowledge that environmental externalities cannot remain afterthoughts. Lifecycle assessments, transparent reporting of energy use, and investment in low-impact alternatives represent essential steps, supported by frameworks such as the AI for Good movement and the UN Sustainable Development Goals (Vinuesa et al. 2020).

To ensure that AI supports long-term public benefit, systems should undergo periodic socio-technical impact assessments that account for human rights, environmental sustainability, and collective well-being. Such evaluations help align innovation with justice, resilience, and responsibility. Key dimensions include:

- **Distributional effects**: Impact assessments should evaluate how the costs and benefits of an AI system were or are distributed across different societal groups (Ren and Wierman 2024; The Indegenous 2025). This includes identifying who benefited or was disadvantaged by the system's outcomes and whether it reduced or amplified existing inequalities. For example, predictive policing or credit scoring tools have historically disadvantaged marginalized populations when designers failed to incorporate explicit equity considerations into system design.
- **Cultural sensitivity**: AI systems must be designed and deployed with awareness of local cultural values, norms, and practices (UNESCO 2023; The Indegenous 2025). This involves avoiding language models or content moderation systems that suppressed Indigenous expressions, misrepresented cultural identities, or ignored linguistic diversity in prior deployments. Respecting cultural plurality remains essential in globally deployed systems such as content recommendation engines or multilingual virtual assistants.
- **Environmental impact**: The environmental footprint of AI, particularly the carbon emissions from training large-scale models or powering real-time inference, has become a central ethical concern (Morrison et al. 2025; Jegham et al. 2025). Developers should measure and mitigate energy usage, promote green computing infrastructures, and adopt efficient architectures whenever possible. As AI continues to scale across domains, its environmental sustainability remains a defining factor in its ethical acceptability.

  - **Public interest alignment**: Evaluations should determine whether AI systems contributed to or undermined public goods such as education, health, safety, accessibility, and democratic governance (Perera et al. 2025). Systems that spread misinformation, reinforced polarization, or reduced civic participation may have achieved technical sophistication yet remained socially harmful. Responsible AI development demands that systems serve collective welfare rather than narrow commercial or political interests.

In summary, societal and environmental well-being has never been a peripheral concern but remains a central pillar of ethical and trustworthy AI. By designing systems that respect ecological boundaries, acknowledge cultural diversity, and promote inclusive prosperity, researchers and practitioners can ensure that technological progress aligns with both human dignity and planetary flourishing.

- **Privacy and Data Governance:**

  Privacy and data governance serve as foundational pillars of trustworthy AI, ensuring that personal and sensitive information is handled ethically, securely, and lawfully throughout the AI system lifecycle (Floridi 2021; Jobin, Ienca, and Vayena 2019). In an increasingly data-driven world, where AI models depend on vast quantities of personal data for training and decision-making, safeguarding individual privacy represents not only a legal requirement but also a moral and social imperative.

  Effective data governance begins with obtaining meaningful and informed user consent, ensuring that individuals understand how their data is collected, used, shared, and stored. This process includes clear communication about data purposes, retention policies, and user rights, such as the ability to access, rectify, or delete personal information. These principles were codified in landmark regulations such as the European Union's General Data Protection Regula-

tion (GDPR), which established the notion of data protection "by design and by default" (Voigt and Von dem Bussche 2017).

To ensure compliance and maintain public trust, AI systems must integrate privacy-enhancing technologies (PETs) such as:

- **Differential Privacy**: This mathematical framework introduces calibrated noise into statistical outputs to protect individual-level information, even when results are aggregated (Dwork 2006; Dwork 2008). It ensures that the inclusion or exclusion of a single record does not significantly affect the overall outcome.
- **Federated Learning**: This decentralized machine learning technique keeps data on local devices while sharing only model updates for aggregation. By minimizing data transfer and central storage, it reduces exposure risks and preserves user privacy (Q. Yang et al. 2019).
- **Homomorphic Encryption and Secure Multiparty Computation**: These cryptographic methods allow AI models to perform computations on encrypted data without accessing raw personal information directly, thereby ensuring confidentiality throughout the computation process (Moriai 2019).

Beyond technical safeguards, data governance demands organizational accountability. It includes implementing robust data access controls, maintaining audit trails, conducting data protection impact assessments (DPIAs), and appointing data protection officers (DPOs) in line with regulatory requirements (European Union 2016). AI developers and institutions must ensure the quality, provenance, and representativeness of their data, as biased or unconsented datasets can produce unfair or harmful outcomes, particularly for vulnerable populations (Mehrabi et al. 2021).

Moreover, privacy in AI extends beyond individual autonomy and intersects with broader societal concerns. Algorithmic systems deployed in policing,

hiring, or credit scoring often draw on behavioral or biometric data without adequate oversight, creating risks of surveillance capitalism or function creep. Privacy-preserving AI must therefore align with democratic norms, human rights, and principles of data justice (Crawford 2021).

In summary, robust privacy and data governance frameworks remain essential for building AI systems that respect human dignity, sustain public trust, and comply with ethical and legal standards. They form the backbone of responsible AI development and deployment across both public and private sectors.

- **Human Agency and Oversight:**

  Human agency and oversight ensure that AI systems serve as instruments of human empowerment rather than mechanisms of automation that erode autonomy, accountability, or moral responsibility. Trustworthy AI requires meaningful human control throughout the system's lifecycle, from data collection and model design to deployment and decision-making (H. AI 2019; Jobin, Ienca, and Vayena 2019; Holstein et al. 2019).

  At the heart of this principle lies the obligation to preserve human dignity and the capacity for both individual and collective decision-making. Instead of delegating complex ethical choices to opaque algorithms, AI should operate as a decision-support tool that complements human expertise and reinforces human values. This requirement is particularly critical in high-stakes contexts such as criminal justice, healthcare, and child welfare, where algorithmic outputs influence life-altering outcomes.

  To uphold human agency, AI systems should incorporate:

  - **Human-in-the-loop (HITL)** mechanisms place humans directly within the decision-making process, ensuring they participate actively in interventions that are automated or partially autonomous (Amershi et al. 2019).

- **Human-on-the-loop** oversight structures enable humans to monitor system outputs, intervene, and override decisions when necessary, particularly in the face of anomalies or ethical dilemmas (H. AI 2019).
- **Human-in-command (HIC)** approaches preserve ultimate human authority over whether, when, and how an AI system operates. They grant decision-makers the discretion to suspend or prohibit deployment in contexts that threaten fundamental rights (H. AI 2019).
- **Informed consent and user control** features empower individuals to understand, contest, or opt out of algorithmic decisions. These mechanisms rely on transparent documentation and user interfaces that support accountability and comprehension (Selbst et al. 2019).

Oversight must extend beyond technical fail-safes to include institutional accountability. Governance bodies conduct algorithmic impact assessments, review bias audits, and establish escalation pathways for appeals and redress. In domains such as predictive policing or automated hiring, human oversight ensures that value-laden decisions are not shaped uncritically by biased data or narrow optimization goals (Raji and Buolamwini 2019).

Furthermore, human agency requires cultivating what the European Commission defines as "AI literacy", the ability of users, developers, and decision-makers to understand, evaluate, and engage critically with AI systems (H. AI 2019). Building AI literacy involves professional training, public awareness initiatives, and the inclusion of diverse stakeholders in the design process, especially those from marginalized communities historically excluded from technological decision-making (Costanza-Chock 2020).

Without sustained human oversight, AI systems risk eroding democratic control, displacing human accountability, and entrenching existing power imbalances. Embedding human agency constitutes not merely a design choice but a democratic and ethical imperative.

- **Consistency:**

  Consistency refers to the stability and reproducibility of an AI system's outputs when it receives identical or semantically equivalent inputs. It represents a key dimension of trustworthy AI by ensuring that the system behaves predictably and avoids arbitrary or erratic decisions under similar conditions. In decision-making contexts, consistency guarantees that individuals in comparable circumstances receive similar treatment, thereby upholding principles of fairness, non-discrimination, and due process (A. Wang, Ramaswamy, and Russakovsky 2022; Madras et al. 2019).

  This property is particularly critical in high-stakes domains such as criminal justice, credit scoring, and healthcare, where inconsistent outputs can produce both perceived and actual injustice. For example, if two loan applicants with nearly identical financial profiles receive different outcomes from a credit model, the system's legitimacy and fairness come into question. In judicial contexts, inconsistent sentencing predictions erode public confidence in automated risk assessment tools and their alignment with legal standards of equity.

  Technically, consistency can be evaluated through:

  - **Input invariance**: The model should produce equivalent outputs when it processes equivalent inputs, even when their representations differ slightly due to encoding, formatting, or rewording (Ribeiro et al. 2020).
  - **Pairwise consistency**: This measure evaluates whether similar individuals receive similar predictions, particularly when minor feature variations or random noise occur in the input data (Dwork et al. 2012).
  - **Longitudinal consistency**: This property ensures that system behavior remains stable over time unless changes result from justified updates in data, context, or policy (Heidari and Krause 2019; Wen, Bastani, and Topcu 2021).

  Although scholars often conflate consistency with reliability, the two represent

distinct concepts. Reliability concerns a model's robustness under varied, real-world conditions, including adversarial or noisy environments, whereas consistency focuses on internal logical coherence across inputs and instances. Both remain essential for trustworthiness: a system that operates reliably but inconsistently may still produce unjust or inexplicable outcomes, while a system that acts consistently but lacks reliability may fail under realistic deployment scenarios. Ensuring consistency also strengthens legal auditability and ethical accountability. In jurisdictions with algorithmic transparency mandates, such as the EU's AI Act and GDPR, consistent system behavior facilitates explanation, contestation, and appeals by affected individuals (European Parliament and Council 2024; B. Goodman and Flaxman 2017).

Ultimately, consistency functions not only as a technical metric but also as a normative commitment to equal treatment. It underpins user trust, safeguards procedural justice, and reinforces the legitimacy of AI systems within society.

2.3.5 From Compliance to Legitimacy

Although ethical guidelines and technical documentation can enhance compliance, trust also depends on legitimacy. A legitimate system is not only lawful but also socially and morally acceptable to the individuals and communities it affects (Binns 2018; Veale, Van Kleek, and Binns 2018). Legitimacy requires participatory governance mechanisms such as stakeholder consultations, community audits, and co-design workshops that incorporate diverse perspectives and values.

For example, fairness interventions may achieve statistical soundness yet lack legitimacy if they fail to reflect the values and lived experiences of affected communities, particularly marginalized groups who bear the brunt of algorithmic harms (Costanza-Chock 2020).

2.3.6 Institutional Frameworks for Trustworthy AI

Several institutional and normative frameworks now guide and standardize the implementation of trustworthy AI principles. Among the most influential are:

- **The EU AI Act (2024)** (European Parliament and Council 2024): This legally binding regulation classifies AI systems by risk level and imposes obligations including transparency, robustness testing, impact assessments, and rights to explanation for high-risk systems.

- **OECD AI Principles** (OECD 2019): These non-binding international guidelines promote human-centric, transparent, fair, and accountable AI, and encourage responsible innovation among member and partner countries.

- **UNESCO Recommendation on the Ethics of Artificial Intelligence** (UNESCO 2021): This global, human rights–centered framework emphasizes multi-stakeholder governance, the protection of fundamental rights, environmental sustainability, and inclusive development.

- **NIST AI Risk Management Framework (AI RMF)** (National Institute of Standards and Technology 2023): This U.S. voluntary framework guides organizations in managing AI risks across lifecycle phases and provides practical tools for governance, risk assessment, monitoring, and assurance.

- **AI Bill of Rights (U.S.)** (White House Office of Science and Technology Policy 2022): This set of principles focuses on safeguarding individual rights in relation to automated systems by promoting fairness, transparency, accountability, and protection against harmful bias.

- **ISO / IEC Standards (e.g., ISO/IEC 42001)** (International Organization for Standardization 2023): These certification-oriented standards establish formal management systems for implementing trustworthy AI practices within organizations.

- **Tools for Transparency: Model Cards and Datasheets** (Mitchell et al. 2019; Gebru et al. 2021): These documentation tools enhance accountability and improve understanding of datasets and AI models, although they are not yet enforceable under law.

These frameworks mark a shift from purely voluntary ethical commitments to enforceable accountability mechanisms. They highlight the growing importance of interdisciplinary governance that brings together ethicists, legal scholars, technologists, and impacted community stakeholders to shape responsible AI development and deployment.

In conclusion, societies do not earn trust in AI through technical performance alone; they build it through transparency, fairness, accountability, and inclusion. Developing trustworthy AI, therefore, requires a sociotechnical perspective that recognizes how algorithmic systems interact with human values, institutional contexts, and historical inequalities. Fairness must reside not only in models and metrics but also in the governance structures and institutional practices that sustain them.

The principle of unfairness in AI systems emerges from biases embedded in data, models, and institutional processes that shape decision-making. These embedded biases serve as the primary mechanisms through which unfairness manifests in practice. The next section examines the sources and forms of algorithmic bias, tracing how these distortions arise and persist across the machine learning pipeline.

2.4 Algorithmic Bias

Algorithmic bias has become a central concern in the deployment of AI systems, especially in high-stakes domains such as criminal justice, healthcare, and finance. Unlike traditional human bias, algorithmic bias operates in more opaque and systemic ways. It arises not only from the datasets that models learn from, but also from the design choices, evaluation criteria, and socio-technical environments in which these systems function.

2.4.1 Sources and Forms of Algorithmic Bias

Bias in machine learning enters the AI pipeline at multiple stages. Scholars typically classify it into three interrelated categories: *data bias*, *model bias*, and *evaluation bias* (Suresh and Guttag 2019; Gu and Oelke 2019; Standardization 2021).

- **Data Bias** emerges from skewed, incomplete, or historically prejudiced training data. In criminal justice, datasets often encode decades of over-policing in minority communities, which inflate recorded crime rates for those groups because of surveillance intensity rather than actual crime prevalence (Obermeyer et al. 2019; Varona and Suárez 2022).

- **Model Bias** occurs when learning algorithms introduce or amplify disparities, even when the data itself is relatively balanced. This bias arises through inappropriate feature selection, over-reliance on correlated but sensitive variables, or flawed regularization schemes that disproportionately affect certain subgroups (D. Kaur et al. 2022; Bechavod and Ligett 2017).

- **Evaluation Bias** results from metrics that fail to capture fairness adequately or from feedback loops that reinforce biased outcomes. For instance, a recidivism prediction tool that optimizes for overall accuracy may ignore disparities in false negative rates across racial groups, thereby producing disproportionate detention recommendations (J. Skeem and C. Lowenkamp 2020; Mishler, E. H. Kennedy, and Chouldechova 2021).

- **Deployment Bias** emerges when technically valid predictions are applied in contexts that amplify or perpetuate inequities. Even a well-calibrated model can produce unfair outcomes if institutions misuse its outputs, disregard uncertainty, or embed predictions within existing systems of disadvantage. For example, a healthcare cost-prediction tool may appear accurate but underallocate resources to minority patients if cost serves as an imperfect proxy for medical need (Obermeyer et al. 2019).

It is also important to distinguish *bias* from *discrimination*. Although scholars often use the terms interchangeably, *bias* describes technical deviations from ideal model behavior, whereas *discrimination* refers to the unjust treatment of individuals based on protected characteristics. Discrimination can manifest as either direct (disparate treatment) or indirect (disparate impact) (Feldman et al. 2015; L. Zhang, Y. Wu, and X. Wu 2017).

Moreover, fairness criteria often conflict in practice. For instance, when base rates differ across groups, it is mathematically impossible for a model to satisfy both calibration and equalized odds simultaneously (Chouldechova 2017; Kleinberg, Mullainathan, and Raghavan 2016). These impossibility results highlight the need to prioritize fairness metrics in a domain-sensitive manner, especially in high-stakes contexts such as recidivism prediction.

2.4.2 Discrimination in Practice: Direct vs. Indirect

In machine learning, **direct discrimination** occurs when sensitive attributes explicitly influence model outcomes. For instance, a sentencing model that includes race as an input variable engages in direct discrimination. In contrast, **indirect discrimination** (or disparate impact) arises when seemingly neutral features serve as proxies for sensitive attributes and produce skewed outcomes. This pattern commonly appears in recidivism prediction tools, where variables such as prior arrests encode institutional bias rather than individual culpability (J. Skeem and C. Lowenkamp 2020; Pin Calmon et al. 2018).

However, not all disparities in model outcomes are ethically or legally equivalent. Some disparities stem from legitimate, task-relevant differences, while others reflect structural or algorithmic injustices. Scholars, therefore, differentiate between two additional categories:

- **Explainable discrimination**: This form of disparity arises when differences in predictions result from valid, non-discriminatory features relevant to the task. For example, variations in loan approval rates may reflect differences

in income levels if income serves as a legitimate indicator of creditworthiness rather than a proxy for a protected attribute. Even in such cases, practitioners must scrutinize features to ensure they do not encode historical inequities or proxy bias.

- **Inexplainable discrimination**: This type of disparity persists after accounting for legitimate explanatory variables, indicating that protected attributes (e.g., race, gender, or age) or their proxies drive outcomes in unjustifiable ways (Varona and Suárez 2022). For instance, when two applicants with similar employment and financial profiles receive systematically different credit decisions based solely on demographic characteristics, the model exhibits ethically and legally problematic discrimination.

Attempts to eliminate discrimination by simply removing protected attributes from training data, an approach known as "fairness through unawareness", prove insufficient because correlated variables often reintroduce the discriminatory signal (Pin Calmon et al. 2018; Jain, Huber, Fegaras, et al. 2019). More promising strategies apply data transformation and probabilistic balancing techniques to suppress discriminatory information while preserving model utility (Pin Calmon et al. 2018).

2.5 Fairness Concepts and Definitions

Fairness remains one of the most contested and multi-dimensional challenges in developing machine learning (ML) systems, particularly in high-stakes domains such as criminal justice, finance, and healthcare. At its core, fairness requires that AI systems avoid systematically disadvantaging individuals or groups based on protected characteristics such as race, gender, age, disability, or socioeconomic status (Barocas, Hardt, and Narayanan 2023; Mehrabi et al. 2021; Caton and Haas 2023). However, translating this normative ideal into computational form poses complex challenges and demands difficult trade-offs.

A foundational distinction in the fairness literature differentiates between two paradigms: **group fairness** and **individual fairness** (Dwork et al. 2012).

- **Group Fairness** aims to ensure equitable outcomes across demographic groups. It draws from legal and policy frameworks that seek to identify and correct systemic disparities. Researchers and practitioners evaluate group fairness using statistical metrics such as demographic parity, disparate impact, equal opportunity, and equalized odds. For example, a recidivism prediction model achieves equal opportunity when it maintains equal true positive rates across racial or gender groups.

- **Individual Fairness** asserts that similar individuals should receive similar treatment. This principle relies on defining a similarity metric based on task-relevant features. For instance, two individuals with comparable criminal histories and socioeconomic backgrounds should obtain equivalent recidivism risk scores, regardless of their gender or race. Although conceptually intuitive, individual fairness remains challenging to implement in practice because defining similarity is subjective and often constrained by limited contextual information (Kusner et al. 2017; Binns 2020).

These two notions of fairness, group and individual fairness, while complementary in principle, often conflict in practice. Optimizing for group-level parity can produce unfair treatment for specific individuals, whereas enforcing individual-level consistency can leave group disparities unresolved. This tension highlights the inherent trade-offs in fairness-aware machine learning.

Given the entrenched and structural nature of disparities in high-stakes applications such as recidivism prediction, this thesis adopts **group fairness** as its primary analytical lens. Group fairness offers direct measurability, aligns with anti-discrimination law and policy, and provides a pragmatic foundation for designing and auditing AI systems in socially sensitive domains (Caton and Haas 2023).

The following section explores how fairness considerations can integrate into the

machine learning pipeline through interventions that span data preparation, model training, and output correction.

2.5.1 Fairness Pipeline Design

To address algorithmic bias, researchers and practitioners apply fairness-enhancing methods at three distinct stages of the machine learning pipeline: pre-processing, in-processing, and post-processing interventions (Friedler et al. 2019). Each stage targets different sources of bias and creates opportunities to mitigate unfair outcomes while balancing predictive performance. This section introduces six widely studied techniques that represent these stages: Reweighing, Disparate Impact Remover, Exponentiated Gradient Reduction, Adversarial Learning, Equalized Odds Optimization, and Reject Option-Based Classification (Bellamy et al. 2019). Together, these methods illustrate the methodological spectrum of fairness interventions commonly examined in the literature (M. Zhang and Sun 2022).

Each technique approaches fairness from a distinct perspective and typically improves at least one common metric, including Statistical Parity Difference (SPD), Disparate Impact (DI), Equal Opportunity Difference (EOD), or Predictive Equality Difference (PED). Pre-processing approaches reduce disparities within datasets, in-processing methods embed fairness constraints during model training, and post-processing techniques adjust model outputs after training to align predictions with fairness goals. No single intervention resolves all fairness challenges comprehensively; instead, the pipeline framework emphasizes how these strategies complement one another across stages of model development.

Pre-processing Methods

- **Reweighing (Re):** Reweighing is a widely adopted pre-processing technique that rebalances the representation of privileged and unprivileged groups during training. The method is based on the intuition that biases in real-world data often arise from disproportionate outcome distributions across groups,

leading machine learning models to learn and amplify existing inequalities. To counter this effect, reweighing assigns instance-specific weights based on both the protected attribute and the class label. It up-weights instances where disadvantaged groups achieve positive outcomes and down-weights those where advantaged groups receive favorable outcomes. In doing so, the method creates an effective training distribution that approximates demographic fairness without altering the original dataset. Reweighing guides the learning algorithm to treat underrepresented patterns as equally important, thereby reducing disparities typically measured by Statistical Parity Difference (SPD) and Disparate Impact (DI) (Kamiran and Calders 2012; Bellamy et al. 2019; Krasanakis et al. 2018; Roh et al. 2021; Jiang and Nachum 2020).

Formally, for a protected attribute $G \in \{\text{privileged (p)}, \text{unprivileged (up)}\}$ and binary outcome $Y \in \{\text{favorable (fav)}, \text{unfavorable (unfav)}\}$, the reweighing algorithm assigns instance weights as:

$$w_{g,y} = \frac{P(G = g)\, P(Y = y)}{P(G = g, Y = y)} \quad \text{for } g \in \{p, up\},\ y \in \{\text{fav}, \text{unfav}\},$$

where probabilities are estimated empirically from the dataset. This adjustment ensures that the joint distribution $P(G, Y)$ factorizes into the marginals $P(G)\, P(Y)$, thereby reducing unwanted dependence between the protected attribute G and the outcome Y.

- **Disparate Impact Remover (DIR):** Disparate Impact Remover focuses on reducing the influence of biased features by transforming the input space itself. The method relies on the intuition that certain features strongly correlate with protected attributes, allowing sensitive information to "leak" into predictions even when the protected attribute is not explicitly included. To counter this, Disparate Impact Remover modifies feature values to weaken these correlations while preserving the relative ordering of individuals to maintain predictive utility. This transformation makes it more difficult for a downstream model

to exploit sensitive signals embedded in correlated variables, thereby reducing disparate impact at the dataset level. The approach demonstrates how fairness can be introduced upstream, before model training, by directly shaping the data used by the algorithm (Feldman et al. 2015; Bellamy et al. 2019).

Formally, let $X = (X_1, \ldots, X_d)$ denote the non-protected features, G the protected attribute, and $\lambda \in [0, 1]$ the repair level. For an individual i in group g with feature value x_{ij}, the repaired value is:

$$x'_{ij} = (1 - \lambda)\, x_{ij} + \lambda\, F_{X_j}^{-1}\big(F_{X_j|G=g}(x_{ij})\big),$$

where $F_{X_j|G=g}$ is the group-conditional cumulative distribution function (CDF) of feature X_j and $F_{X_j}^{-1}$ is the inverse CDF of the pooled distribution across all groups. Setting $\lambda = 0$ leaves the data unchanged, while $\lambda = 1$ fully repairs features by aligning group distributions to the pooled marginal.

In-processing Methods

- **Exponentiated Gradient Reduction (EGR):** Exponentiated Gradient Reduction is an in-processing method that frames fair classification as an optimization problem with explicit fairness constraints. The approach treats fairness not as a post hoc adjustment but as a requirement embedded directly into the model's objective function. It decomposes the learning task into a sequence of cost-sensitive classification problems, where each iteration balances prediction accuracy against fairness criteria. Using an exponentiated gradient algorithm, the method iteratively adjusts the classifier to minimize empirical error while satisfying fairness constraints such as demographic parity, equal opportunity, and predictive equality. In doing so, Exponentiated Gradient Reduction offers a principled and generalizable framework for training models that integrate fairness objectives directly into the learning process (Agarwal et al. 2018; Bellamy et al. 2019).

Formally, the method learns a randomized classifier Q by solving

$$\min_{Q \in \Delta(\mathcal{H})} \mathbb{E}_{h \sim Q}\left[\frac{1}{n}\sum_{i=1}^{n} \ell(h(x_i), y_i)\right] \quad \text{s.t.} \quad Q \in \mathcal{C}_\varepsilon,$$

where $\Delta(\mathcal{H})$ is the set of distributions over base classifiers, ℓ is the loss (e.g., 0–1 loss), and $\mathcal{C}_\varepsilon$ encodes the fairness constraints (e.g., Demographic Parity or Equalized Odds) within tolerance ε. The solution is obtained via exponentiated gradient updates on dual variables, returning a mixture of classifiers that satisfies fairness requirements while maintaining low error.

- **Adversarial Learning (AL):** Adversarial Learning functions as an in-processing strategy that applies adversarial training principles to promote fairness. The approach rests on the intuition that if a model's predictions encode information about protected attributes, an adversary should recover those attributes from the predictions. To counter this, Adversarial Learning establishes a two-player game between a predictor and an adversary: the predictor maximizes task performance, while the adversary attempts to infer protected attributes from the predictor's outputs. The predictor receives a penalty whenever the adversary succeeds, which incentivizes it to remove sensitive information from its internal representations. This adversarial dynamic drives the model to reduce its reliance on protected characteristics, even indirectly through correlated features. Empirical studies show that Adversarial Learning effectively mitigates indirect bias and improves fairness metrics such as Statistical Parity Difference, Equal Opportunity Difference, and Disparate Impact (B. H. Zhang, Lemoine, and Mitchell 2018; Bellamy et al. 2019; Wadsworth, Vera, and Piech 2018; Beutel et al. 2017; Edwards and Storkey 2015).

  Formally, let h_θ denote the predictor with parameters θ and a_ϕ the adversary with parameters ϕ. The predictor loss is

$$L_C(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell\big(h_\theta(x_i), y_i\big),$$

and the adversary loss is

$$L_A(\theta, \phi) = \frac{1}{n} \sum_{i=1}^{n} \ell\big(a_\phi(u_i), g_i\big),$$

where ℓ is binary cross-entropy, y_i are true labels, g_i are protected attributes, and $u_i = [\sigma(s_i), \sigma(s_i)\cdot y_i, \sigma(s_i)\cdot(1-y_i)]$ are the enriched inputs to the adversary with $s_i = h_\theta(x_i)$ the predictor logits. The classifier parameters are updated via

$$\theta \leftarrow \theta - \eta \left[\nabla_\theta L_C - \mathrm{proj}_{\hat{g}_A}(\nabla_\theta L_C) - \lambda \nabla_\theta L_A\right],$$

where $\hat{g}_A = \nabla_\theta L_A / \|\nabla_\theta L_A\|$ is the normalized adversary gradient, $\mathrm{proj}_u(v) = \frac{v^\top u}{\|u\|^2} u$ is the projection operator, and $\lambda > 0$ controls the adversarial penalty. The adversary parameters ϕ are updated to minimize L_A, thereby improving its predictive accuracy. At equilibrium, the predictor achieves low error while producing outputs that are approximately independent of the protected attribute.

Post-processing Methods

- **Reject Option-Based Classification (ROC):** Reject Option-Based Classification serves as a post-processing method that modifies model predictions in regions of uncertainty, particularly near the decision boundary. The method is based on the intuition that when a model exhibits low confidence in its predictions, this uncertainty presents an opportunity to adjust decisions in favor of fairness. In implementation, the algorithm assigns favorable outcomes to members of unprivileged groups and assigns unfavorable outcomes to members of privileged groups within this uncertainty region. This targeted reassignment preserves the underlying model while strategically altering its outputs to mitigate disparities. Consequently, Reject Option-Based Classification reduces bias across several fairness metrics, including Statistical Parity Difference, Disparate Impact, Equal Opportunity Difference, and Predictive

Equality Difference (Kamiran, Karim, and X. Zhang 2012; Bellamy et al. 2019; Gao et al. 2022).

Formally, let $\hat{s}_i \in [0,1]$ be the predicted score for instance i, τ a classification threshold, and Δ a margin around the threshold. The base decision rule is

$$\hat{y}_i = \begin{cases} 1, & \hat{s}_i > \tau, \\ 0, & \hat{s}_i \leq \tau, \end{cases}$$

and the *critical region* of uncertainty is defined as

$$\mathcal{C} = \{\, i : \tau - \Delta < \hat{s}_i \leq \tau + \Delta \,\}.$$

For individuals in $\mathcal{C}$, predictions are adjusted as

$$\hat{y}_i = \begin{cases} 1, & i \in \mathcal{C},\ g_i = \text{unprivileged}, \\ 0, & i \in \mathcal{C},\ g_i = \text{privileged}, \\ \text{original rule}, & i \notin \mathcal{C}, \end{cases}$$

where g_i denotes the protected group membership. The optimal pair (τ^*, Δ^*) is chosen to maximize balanced accuracy subject to fairness constraints on the chosen metric.

- **Equalized Odds Optimization (EO):** Equalized Odds Optimization directly addresses disparities in error rates across demographic groups. The approach rests on the intuition that fairness requires not only balanced positive outcomes but also equal predictive quality across protected categories. It applies linear programming to adjust predicted labels or output probabilities so that both true positive rates and false positive rates align between groups. By modifying predictions after model training, Equalized Odds Optimization ensures that individuals from different groups receive similar treatment in

terms of classification errors. The method proves particularly effective for fairness metrics such as Equal Opportunity Difference and Predictive Equality Difference, providing a mathematically principled way to align predictive performance with fairness criteria (Hardt, Price, and Srebro 2016; Pleiss et al. 2017; Bellamy et al. 2019).

Formally, let $\hat{y}$ be the model's predicted label and $g \in \{0, 1\}$ denote the protected attribute (0 = privileged, 1 = unprivileged). The postprocessing introduces randomized relabeling probabilities

$$\alpha_g = \Pr(\tilde{y} = 1 \mid \hat{y} = 1, G = g), \qquad \beta_g = \Pr(\tilde{y} = 1 \mid \hat{y} = 0, G = g),$$

so that the final prediction $\tilde{y}$ is drawn from these distributions. The adjusted true- and false-positive rates are

$$\mathrm{TPR}'_g = \alpha_g \,\mathrm{TPR}_g + \beta_g \,(1 - \mathrm{TPR}_g), \qquad \mathrm{FPR}'_g = \alpha_g \,\mathrm{FPR}_g + \beta_g \,(1 - \mathrm{FPR}_g),$$

where TPR_g and FPR_g are the original groupwise rates. The optimization problem solved is

$$\begin{aligned} \min_{\alpha_0,\beta_0,\alpha_1,\beta_1} \quad & c^\top x \\ \text{s.t.} \quad & \mathrm{TPR}'_0 = \mathrm{TPR}'_1, \\ & \mathrm{FPR}'_0 = \mathrm{FPR}'_1, \\ & 0 \le \alpha_g, \beta_g \le 1 \quad (g = 0, 1), \end{aligned}$$

where $x = (\alpha_0, \beta_0, \alpha_1, \beta_1)^\top$ and c is a vector of coefficients derived from the original error rates. The optimal solution provides group-specific flipping probabilities that enforce equalized odds while preserving as much predictive accuracy as possible.

2.6 Fairness and Performance Metrics

This section outlines the concept of fairness in AI and examines contemporary methodologies that evaluate and enhance it, particularly within binary classification tasks such as recidivism prediction. It focuses on group fairness, employing both aggregate and subgroup-level metrics to assess equitable model behavior.

2.6.1 Fairness Metrics

This study employs several group fairness metrics, evaluated at both the aggregate level (privileged vs. all unprivileged groups) and subgroup level (each unprivileged subgroup vs. the privileged group). Let $\hat{y} \in \{0, 1\}$ denote the predicted label, $y \in \{0, 1\}$ the true label, and G the protected group variable. Let $G = g_{\text{ref}}$ be the privileged group, and $G = g_k$ be an unprivileged subgroup.

- **Statistical Parity Difference (SPD):** Statistical Parity Difference (SPD) quantifies whether individuals from different demographic groups are equally likely to receive a favorable prediction (i.e., $\hat{y} = 1$), regardless of their true outcome (y). This metric does not consider actual labels, making it a measure of representational parity. A value of zero indicates that the likelihood of receiving a favorable outcome is equal across groups, suggesting fair treatment at the group level. Positive or negative values signal potential disparities in how the model treats privileged and unprivileged groups.

  For example, in a recidivism prediction task, assume the model assigns a "high risk" label to 60% of Black defendants and 40% of White defendants. Then:

$$SPD = P(\hat{y} = 1 \mid G = \text{Black}) - P(\hat{y} = 1 \mid G = \text{White}) = 0.6 - 0.4 = 0.2$$

  An SPD of 0.2 means that Black individuals are 20 percentage points more likely to be labeled as high risk than White individuals, indicating potential

bias in the model's predictions.

- *Aggregate SPD:*

$$SPD_{\text{agg}} = P(\hat{y} = 1 \mid G = g_{\text{ref}}) - P(\hat{y} = 1 \mid G \in \text{Unprivileged Groups})$$

- *Subgroup SPD (for each g_k):*

$$SPD_{g_k} = P(\hat{y} = 1 \mid G = g_{\text{ref}}) - P(\hat{y} = 1 \mid G = g_k)$$

- **Disparate Impact (DI):** Disparate Impact (DI) assesses fairness by examining the ratio of favorable predictions ($\hat{y} = 1$) between unprivileged and privileged groups. It is widely used in legal and regulatory contexts, especially under the U.S. Equal Employment Opportunity Commission's 80% rule, which considers DI values below 0.8 as indicative of potential discrimination. A DI value of 1 implies equal treatment across groups, while deviations from 1 reflect disproportional benefits or disadvantages in predictions.

  For instance, suppose that in a credit approval scenario, 60% of Black applicants receive loan approvals ($P(\hat{y} = 1 \mid G = \text{Black}) = 0.6$), while 80% of White applicants receive the same outcome ($P(\hat{y} = 1 \mid G = \text{White}) = 0.8$). Then:

$$DI = \frac{0.6}{0.8} = 0.75$$

  A DI of 0.75 suggests that Black applicants are 25% less likely to be approved for a loan compared to White applicants, potentially flagging unfair treatment and raising legal concerns.

  - *Aggregate DI:*

$$DI_{\text{agg}} = \frac{P(\hat{y} = 1 \mid G \in \text{Unprivileged Groups})}{P(\hat{y} = 1 \mid G = g_{\text{ref}})}$$

– *Subgroup DI (for each g_k):*

$$DI_{g_k} = \frac{P(\hat{y} = 1 \mid G = g_k)}{P(\hat{y} = 1 \mid G = g_{\text{ref}})}$$

- **Equal Opportunity Difference (EOD):** Equal Opportunity Difference evaluates whether individuals who genuinely qualify for a favorable outcome (i.e., those with $y = 1$) receive equal treatment across demographic groups. Specifically, it measures the difference in the true positive rate (TPR) between the privileged and unprivileged groups. An EOD value of zero indicates that all qualified individuals, regardless of group membership, are equally likely to receive a positive prediction from the model, thus satisfying the fairness criterion.

  For example, suppose that among those who are truly eligible for parole (i.e., $y = 1$), 75% of Black individuals are correctly predicted as low-risk ($P(\hat{y} = 1 \mid y = 1, G = \text{Black}) = 0.75$), while only 65% of White individuals receive the same correct prediction ($P(\hat{y} = 1 \mid y = 1, G = \text{White}) = 0.65$). The EOD in this case would be:

$$EOD = 0.65 - 0.75 = -0.10$$

  A negative EOD value indicates that the unprivileged group (Black individuals, in this case) is more likely to receive favorable outcomes among those who qualify, which may imply biased advantages, or the inverse, if the values are reversed.

– *Aggregate EOD:*

$$EOD_{\text{agg}} = P(\hat{y} = 1 \mid y = 1, G = g_{\text{ref}}) - P(\hat{y} = 1 \mid y = 1, G \in \text{Unprivileged Groups})$$

– *Subgroup EOD (for each g_k):*

$$EOD_{g_k} = P(\hat{y} = 1 \mid y = 1, G = g_{\text{ref}}) - P(\hat{y} = 1 \mid y = 1, G = g_k)$$

- **Predictive Equality Difference (PED):** Predictive Equality Difference (PED) quantifies the disparity in false positive rates (FPR) between the privileged and unprivileged groups. A false positive occurs when the model incorrectly assigns a positive prediction to someone who should have received a negative one (i.e., $y = 0$, but $\hat{y} = 1$). A PED value of zero reflects fairness, as it indicates that all groups are equally likely to be incorrectly labeled as high risk or eligible.

  Disparities in PED are particularly critical in high-stakes settings like criminal justice or healthcare, where a false positive can lead to unjust consequences, such as unnecessary incarceration or inappropriate treatment.

  For example, suppose that among individuals who are not expected to reoffend ($y = 0$), 45% of Black individuals are incorrectly predicted as high-risk ($\hat{y} = 1$), while only 25% of White individuals receive the same incorrect prediction. The PED in this case is:

$$PED = 0.25 - 0.45 = -0.20$$

  A negative PED value like this suggests that the unprivileged group (Black individuals) faces a disproportionately higher risk of false positive predictions, raising fairness concerns about disparate mistreatment.

– *Aggregate PED:*

$$PED_{\text{agg}} = P(\hat{y} = 1 \mid y = 0, G = g_{\text{ref}}) - P(\hat{y} = 1 \mid y = 0, G \in \text{Unprivileged Groups})$$

– *Subgroup PED (for each g_k):*

$$PED_{g_k} = P(\hat{y} = 1 \mid y = 0, G = g_{\text{ref}}) - P(\hat{y} = 1 \mid y = 0, G = g_k)$$

- **Accuracy:** Accuracy refers to the proportion of total predictions that the model classifies correctly. It is one of the most widely used performance metrics in machine learning because it captures overall predictive success. However, while aggregate accuracy provides a general sense of model effectiveness, it may mask important disparities in how well the model performs for different demographic groups.

  For instance, a model might achieve high overall accuracy by performing very well on the majority (privileged) group while underperforming on minority (unprivileged) groups. This creates an illusion of fairness, even when significant group-level harms exist. To address this, subgroup accuracy is also reported to uncover such disparities.

  Subgroup accuracy reveals how often predictions are correct within specific protected groups (e.g., race or gender subpopulations). This disaggregated view is crucial for identifying and correcting model behavior that may disadvantage particular demographics.

  – *Aggregate Accuracy:*

$$\text{Accuracy}_{\text{agg}} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

  – *Subgroup Accuracy (for each g_k):*

$$\text{Accuracy}_{g_k} = \frac{\text{Correct Predictions in Subgroup } g_k}{\text{Total Samples in Subgroup } g_k}$$

2.6.2 Fairness-Metric Trade-offs

Achieving fairness in machine learning remains inherently complex because commonly used fairness metrics often conflict with one another (Kleinberg, Mullainathan, and Raghavan 2016; Chouldechova 2017; Z. Chen et al. 2023). Each metric reflects a distinct normative or statistical interpretation of equity, and optimizing one metric frequently deteriorates another. Scholars refer to this tension as the "fairness impossibility problem," highlighting that it is mathematically infeasible to satisfy all fairness criteria simultaneously. The following summarizes some of the most frequently discussed trade-offs in the literature.

- **Statistical Parity vs. Equal Opportunity:** Statistical Parity seeks to equalize the overall rate of favorable outcomes across groups, regardless of differences in ground truth. While this approach promotes representational balance, it can inadvertently advantage or disadvantage individuals relative to their qualifications. This often conflicts with Equal Opportunity, which requires that equally qualified individuals across groups have an equal chance of achieving a positive outcome.

- **Equal Opportunity vs. Predictive Equality:** Equal Opportunity ensures that true positive rates align across groups so that qualified individuals receive comparable treatment. However, this focus overlooks disparities in false positives, which can produce unequal misclassification rates. Predictive Equality, by contrast, demands equal false positive rates but may sacrifice parity in true positives. These criteria, therefore, impose opposing constraints on model performance.

- **Disparate Impact vs. Equal Opportunity:** Disparate Impact, often measured using the 80% rule, requires parity in group-level outcome rates. Although this metric helps identify systemic disadvantage, it does not account for individual merit or qualification. As a result, optimizing for proportional group ratios can conflict with Equal Opportunity by allocating favorable out-

comes to less-qualified individuals solely to satisfy statistical parity thresholds.

In summary, these trade-offs demonstrate that fairness is not a single, universally definable construct but rather a set of competing objectives. This study, therefore, adopts a multi-metric, multi-stage framework to balance these objectives and evaluate fairness from complementary perspectives.

These inherent incompatibilities highlight the need for methodological approaches that systematically navigate fairness trade-offs. This thesis addresses that challenge by employing multi-objective optimization (Chapter 4), which enables the principled exploration of fairness–accuracy trade-offs across multiple, potentially conflicting metrics.

2.6.3 Ripple Effects in Applied Contexts

While the previous subsection outlined the theoretical conflicts among fairness metrics, these tensions gain greater significance when models operate in real-world, high-stakes contexts. Optimizing for one fairness dimension often produces unintended "ripple effects" across others, thereby amplifying the ethical and practical dilemmas that decision-makers face.

For example, enforcing *Statistical Parity* in recidivism prediction ensures that demographic groups receive favorable classifications at equal rates. However, this approach may misalign individual treatment, favoring some less-qualified candidates while penalizing qualified ones. Such outcomes undermine *Equal Opportunity*, which emphasizes merit-based fairness. Conversely, improving *Equal Opportunity Difference* by equalizing true positive rates does not address false positives, which can increase disparities in *Predictive Equality Difference*. In sensitive domains such as criminal justice or healthcare, these disparities carry serious consequences: false positives may lead to unjust incarceration or unnecessary treatment.

A system can also satisfy the legal threshold for *Disparate Impact* while concealing inequities in error distribution or access to opportunities. In such cases, a model may appear compliant by one measure yet perpetuate structural disadvantage in

practice. This demonstrates the inadequacy of single-metric evaluation as a proxy for genuine fairness.

These ripple effects underscore the need for a holistic, multi-metric evaluation strategy. By explicitly accounting for competing objectives and domain-specific priorities, and by leveraging multi-objective optimization techniques, practitioners can navigate fairness trade-offs more effectively. Only through such a comprehensive evaluation can fairness interventions achieve both technical rigor and social relevance.

2.6.4 Limitations and Normative Tensions

Fairness metrics, though technically precise, often oversimplify the structural and historical inequalities embedded in real-world data. For instance, a model may satisfy statistical parity yet still encode discriminatory patterns when the underlying data reflect biased institutional practices (Eubanks 2018; Birhane 2021). Moreover, most fairness metrics presume binary group categories and overlook intersectional identities, thereby leaving compounded disadvantages unaddressed.

Another key limitation lies in the normative nature of fairness itself. Stakeholders across and within domains frequently disagree on what constitutes fair treatment. A fairness criterion that appears acceptable in credit lending may prove ethically problematic in criminal justice. These divergent perspectives underscore the need for a comprehensive evaluation of fairness interventions across the entire ML pipeline, which can help reconcile competing notions of fairness in practice. Consequently, the selection of a fairness definition cannot remain a purely mathematical exercise; it must also reflect domain-specific requirements, legal standards, and societal values (Barocas, Hardt, and Narayanan 2023; Binns 2018).

2.6.5 Fairness as a Sociotechnical Construct

Recent literature emphasizes that fairness should not be understood as a property of algorithms alone but as a feature of the broader sociotechnical systems in which they

operate (Selbst et al. 2019; Veale, Van Kleek, and Binns 2018). This perspective highlights that model design, data collection, institutional practices, and power dynamics jointly shape algorithmic outcomes.

From this standpoint, fairness functions as an ongoing process rather than a static goal. It requires continuous auditing, stakeholder participation, and ethical reflection to remain contextually relevant and socially responsive. In high-stakes domains such as recidivism prediction, fairness must account for historical injustice, lived experience, and the intersecting dimensions of identity and marginalization that influence both data and decision-making.

2.7 Intersectionality in Algorithmic Fairness

Traditional fairness approaches often focus on single protected attributes such as race or gender. However, individuals occupy multiple social identities, including class, disability, and immigration status, that interact to shape unique experiences of privilege and marginalization. The theory of *intersectionality*, first introduced by Crenshaw (Crenshaw 2022), explains how overlapping social identities produce compounded and context-dependent forms of disadvantage. Consequently, single-axis fairness assessments risk overlooking harms that emerge within intersectional subgroups.

For example, a predictive model may appear unbiased when evaluated separately across race and gender, yet still systematically disadvantage Black women once subgroup interactions are considered. Kearns et al. (Kearns et al. 2018) describe this phenomenon as "fairness gerrymandering," in which algorithms satisfy aggregate fairness criteria while perpetuating inequities for smaller or intersecting groups. Without intersectional analysis, algorithmic systems risk reinforcing the very disparities they are designed to mitigate (Gohar and L. Cheng 2023).

Intersectionality extends beyond technical measurement to represent both an ethical and epistemic imperative in the design and governance of AI systems (Birhane 2021; Gohar and L. Cheng 2023). Populations such as transgender women of color,

disabled immigrants, or low-income Black mothers experience overlapping disadvantages that remain invisible within one-dimensional fairness frameworks. Embedding intersectionality in algorithmic fairness requires designers, auditors, and policymakers to move beyond abstract parity metrics and ask explicitly: *Who is most at risk, and how can their protection be prioritized?* This reframing positions fairness as a justice-oriented practice rooted in multidimensional social realities rather than as a purely statistical or procedural exercise.

In summary, fairness in machine learning remains a contested and multifaceted concept. While group and individual fairness metrics offer valuable technical tools, they cannot capture the full complexity of structural inequality. A fairness-aware pipeline must therefore integrate a sociotechnical perspective that acknowledges institutional context, normative tensions, and intersectional harms. Only through such a multidimensional approach can AI systems avoid perpetuating injustice and contribute to more equitable and trustworthy outcomes.

2.8 Conclusion

This chapter provides the conceptual and definitional foundations that inform the rest of this thesis. It situates fairness within the broader framework of Trustworthy AI, where principles such as transparency, robustness, accountability, and human oversight interact. Among these dimensions, fairness emerges as particularly critical in high-stakes decision-making domains because it directly influences the distribution of opportunities and harms.

The chapter examines how algorithmic bias arises through data, labels, models, and deployment contexts, illustrating both direct and indirect forms of discrimination. It builds on this foundation to review key fairness concepts and definitions, distinguishing between group fairness, individual fairness, and intersectional fairness. The discussion highlights the incompatibility of certain fairness criteria, underscoring the need for careful metric selection and acknowledging that not all fairness goals can be achieved simultaneously.

The chapter also introduces fairness and performance metrics in detail, emphasizing their complementary perspectives on allocation and error distribution. It identifies the fairness–accuracy trade-off as a recurring challenge with implications for both theory and practice. This tension connects to optimization-based approaches, which offer principled methods for navigating competing objectives. The analysis foregrounds intersectionality as both an ethical and methodological priority, ensuring that evaluations do not obscure compounded biases across demographic subgroups.

Taken together, this chapter establishes fairness not only as a technical requirement but also as a sociotechnical construct shaped by values, institutions, and context. It highlights current limitations in fairness-aware AI research, including fragmented interventions, unresolved trade-offs, and limited generalizability across datasets and domains. These gaps motivate the integrated methodological framework developed in Chapter 4, where fairness interventions are systematically combined across pipeline stages and evaluated through optimization and statistical significance testing.

Chapter 3

Literature Review: Trustworthy AI, Algorithmic Fairness, and Intersectionality

3.1 Introduction

This chapter reviews the literature on trustworthy artificial intelligence (AI), emphasizing algorithmic fairness, bias mitigation, and intersectionality in high-stakes decision-making domains. Building on the conceptual foundations introduced in Chapter 2, it moves beyond definitions to critically synthesize technical, regulatory, and ethical perspectives, highlighting points of convergence and contradiction.

It situates the researcher's prior work on fairness-aware machine learning, intersectional auditing, and empirical evaluation in high-stakes domains within the broader evolution of literature and policy on trustworthy AI. The discussion draws together insights from systematic reviews, methodological innovations, and applied studies, along with recent developments such as the European Commission's Ethics Guidelines for Trustworthy AI (H. AI 2019), advances in fairness-aware learning techniques, and emerging domain-specific auditing practices. In doing so, it bridges earlier debates on algorithmic fairness with contemporary approaches to socio-technical

accountability.

The chapter is structured as follows. Section 3.2 outlines and reviews the foundations of trustworthy AI. Sections 3.3 and 3.4 examine fairness definitions and mitigation strategies, while Section 3.5 explores intersectional perspectives on AI in decision-making. Section 3.6 discusses the limitations of isolated interventions and the need for integrated frameworks. Section 3.7 synthesizes persistent challenges, and Section 3.8 concludes with a transition to the methodology chapter.

3.2 Ethics and Trustworthy AI

The concept of trustworthy artificial intelligence (AI) has become a central concern in debates about the responsible design and deployment of AI systems (H. Liu et al. 2022). Early frameworks, such as those proposed by the European Commission's High-Level Expert Group on AI (H. AI 2019), crystallized key principles including fairness, accountability, transparency, privacy, robustness, and human oversight. Professional and intergovernmental organizations such as IEEE (Tong et al. 2025), UNESCO (UNESCO 2021), and OECD (OECD 2019) have articulated similar principles, reflecting an emerging global consensus on the need to balance innovation with the protection of fundamental rights. While these frameworks established an essential normative foundation, critics argue that they remain overly abstract, offering few concrete mechanisms for embedding principles such as fairness into the socio-technical realities of algorithmic systems (Mittelstadt 2019; Birhane 2021). Fairness, in particular, has emerged as both the most contested and the most operationalized principle, especially in high-stakes domains such as criminal justice, where algorithmic risk assessment tools shape decisions on sentencing, parole, and bail.

The trajectory of research and practice has gradually shifted from aspirational principles toward implementation frameworks. Scholars emphasize the role of algorithmic auditing in uncovering systemic inequities in commercial AI systems (Raji, Smart, et al. 2020) and highlight the centrality of interpretability for accountability,

underscoring that explanations are necessary for evaluating fairness claims (Doshi-Velez and B. Kim 2017). Regulatory initiatives further institutionalize these efforts: the EU AI Act (2024) represents the first binding legal framework to codify obligations around fairness, transparency, and accountability, while the NIST AI Risk Management Framework (2023) provides structured guidance for organizations to embed trustworthiness throughout the AI lifecycle. These developments mark a significant shift, demonstrating how fairness has moved from an aspirational ideal to a legal and organizational requirement, while simultaneously revealing the practical difficulties of operationalizing fairness in complex systems.

Despite this progress, scholars caution against equating fairness with technical compliance. Birhane (Birhane 2021) critiques the "algorithmic justice gap," arguing that fairness metrics alone cannot address the more profound structural inequities that shape algorithmic outcomes. Selbst et al. (Selbst et al. 2019) warn of "fairness formalism," whereby ethical concerns collapse into narrow mathematical criteria detached from social context. Governance research points to the risk of "ethics washing," whereby voluntary guidelines lack enforcement and serve primarily symbolic functions (Bietti 2020). Decolonial perspectives extend these critiques, insisting that trustworthy AI must adopt justice-oriented approaches attentive to power asymmetries, particularly in high-stakes applications such as criminal justice (Mhlambi and Tiribelli 2023). These critiques resonate strongly in the domain of recidivism prediction, where scholars highlight the absence of a consensus definition of fairness (R. Berk, Heidari, Jabbari, Kearns, et al. 2021; O'Loughlin and Bukowitz 2021) and demonstrate that public perceptions of fairness diverge depending on whether stakeholders prioritize reliability, relevance, or causal explanations of disparities (Grgic-Hlaca et al. 2018).

Recent research has expanded the scope of trustworthiness to address new technological and regulatory challenges. Weidinger et al. (Weidinger et al. 2022) identify risks associated with large language models, including misinformation, representational harms, and environmental costs, while Bommasani et al. (Bommasani 2021)

highlight governance issues such as opacity and concentration of power within a small number of organizations. Data governance has similarly been foregrounded as a means of reinforcing fairness and accountability, with dataset documentation and lineage tracking proposed as mechanisms for improving transparency in machine learning pipelines (Stoyanovich, Howe, and Jagadish 2020). Within criminal justice, transparency has repeatedly been emphasized as a precondition for fairness and trust. Public reports document the opacity of proprietary risk assessment tools and their lack of independent scrutiny (McKay 2020; Lo Piano 2020; Chugh 2021). Yet, full transparency remains contested due to privacy risks and the commercial interests of vendors (Caroline Wang et al. 2022).

A central insight emerging from these literatures is that fairness, and by extension, trustworthiness, must be understood as socio-technical rather than purely technical. Trust is relational, shaped not only by algorithms but also by the practices of regulators, developers, institutions, and the communities most affected. Studies show that stakeholders place greater weight on concrete fairness and accountability mechanisms than on abstract principles (Chiou and J. D. Lee 2023; Sousa et al. 2024). Participatory approaches demonstrate that involving affected communities in defining fairness and oversight criteria strengthens legitimacy and reflexivity in governance (Delgado et al. 2023). In criminal justice specifically, fairness and accountability for algorithmic harms remain weak, with responsibility often diffused across developers, private vendors, and public authorities (Hartmann and Wenzelburger 2021; Figueroa-Armijos, B. B. Clark, and Motta Veiga 2022). Tools are frequently deployed without meaningful engagement of frontline justice officials or impacted communities, raising profound concerns about legitimacy and fairness (Cadigan and C. T. Lowenkamp 2011; Mökander et al. 2022).

In summary, fairness emerges from the discourse on trustworthy AI as both the most contested and the most critical principle for ensuring legitimacy in high-stakes contexts. Nowhere is this more evident than in recidivism prediction, where fairness functions as the hinge between technical design and normative judgment.

The following section, therefore, examines algorithmic fairness in greater detail, analyzing competing definitions, trade-offs, and technical debates that structure this thesis.

3.3 Algorithmic Fairness Related Work

Building on the conceptual definitions of group and individual fairness introduced in Chapter 2, this section examines how the literature has extended and critiqued these frameworks in practice. Rather than reiterating definitions, it highlights recent developments that expand fairness toward intersectional, multi-objective, and socio-technical perspectives.

Beyond group and individual fairness, a growing body of research investigates compounded harms that emerge at the intersections of multiple attributes. Kearns et al. (Kearns et al. 2018) introduced the notion of fairness gerrymandering, showing that models can appear fair across single attributes while systematically disadvantaging intersectional subgroups, such as Black women. Extensions including worst-case subgroup fairness (Ghosh, Genuit, and Reagan 2021), differential fairness (Foulds et al. 2020), and multicalibration (Hébert-Johnson et al. 2018) seek to extend fairness guarantees across a richer space of subgroups. In recidivism prediction, subgroup-sensitive interventions such as Protected-Category SMOTE and FAWOS target compounded disparities (Popoola and J. Sheppard 2024; Salazar et al. 2021). Empirical studies across healthcare and finance confirm that neglecting intersections of attributes like race, gender, and socioeconomic status produces systematic disadvantages (Koçak et al. 2025; Valentine, Charney, and Landi 2024; Morina et al. 2019). Embedding subgroup auditing into integrated fairness pipelines, therefore, offers a more accurate account of interventions' impact (Kearns et al. 2018; M. P. Kim, Ghorbani, and J. Zou 2019).

A central theme in this literature is that fairness definitions are not only contested but also mathematically incompatible. Predictive parity, calibration, and equalized odds cannot all be satisfied when base rates differ (Kleinberg, Mullainathan, and

Raghavan 2016; Chouldechova 2017). Empirical studies demonstrate that enforcing calibration exacerbates disparities in error rates, while constraining error rates undermines calibration (Pleiss et al. 2017). Oversampling methods illustrate these tensions: some approaches improve parity but reduce accuracy, while subgroup-sensitive techniques mitigate disparities yet raise concerns about robustness and interpretability (Zelaya 2019; Hickman et al. 2023). Consequently, multi-objective optimization reframes these trade-offs by treating fairness and accuracy as competing objectives and locating non-dominated solutions along Pareto frontiers (G. Yu et al. 2025; Nagpal et al. 2024).

Frameworks and Surveys. Scholars have sought to synthesize these diverse approaches through conceptual frameworks and surveys. Mehrabi et al. (Mehrabi et al. 2021) and Caton et al. (Caton and Haas 2024) propose comprehensive taxonomies, while the Fairness Tree maps relationships among fairness metrics to guide practitioners (Castelnovo et al. 2022). Sectoral reviews extend these efforts into applied domains, such as healthcare (Rajkomar et al. 2018) and fairness benchmarking (Le Quy et al. 2022). However, critics argue that many frameworks remain overly technical, neglecting the historical, institutional, and structural contexts that shape fairness (Selbst et al. 2019; Mhlambi and Tiribelli 2023; Birhane 2020). To address these gaps, recent work emphasizes the integration of ethical reflection and participatory governance with statistical fairness, situating fairness within broader notions of trustworthy AI (McCormack and Bendechache 2024; H. AI 2019; Magaña and Shilton 2025).

In summary, algorithmic fairness encompasses a diverse repertoire of definitions and frameworks, each capturing partial and often conflicting dimensions of equity. These conceptual tensions underscore the difficulty of translating fairness ideals into practice and highlight the need for concrete interventions that operationalize fairness across the machine learning pipeline. The following section, therefore, turns to bias mitigation strategies, examining approaches at the pre-processing, in-processing, and post-processing stages.

3.4 Bias Mitigation Strategies Across the Machine Learning Pipeline

Building on the definitional and conceptual foundations of fairness outlined in Chapter 2, this section examines how fairness is operationalized through bias mitigation strategies across the machine learning (ML) pipeline. While formal definitions provide evaluative criteria, a parallel body of research investigates practical interventions that reduce bias in data, algorithms, and outcomes. These methods are conventionally grouped into pre-processing, in-processing, and post-processing approaches, each targeting different stages of the ML lifecycle. Collectively, they illustrate both the technical possibilities and the limitations of implementing fairness in practice.

Fairness remains the most scrutinized requirement of trustworthy AI, particularly when applying predictive models to assess recidivism risk. Accordingly, this section enumerates and discusses the state-of-the-art techniques that promote fairness and accountability in AI models within the recidivism context. These interventions are analyzed in relation to the stages of AI development identified in the literature and in Section 2.5.1. As previously established, three broad phases structure fairness implementation within ML pipelines: the pre-processing, in-processing, and post-processing phases. This thesis maps and analyzes the leading methodologies across these phases and synthesizes their implications for the design of fairness-aware systems in high-stakes domains.

3.4.1 Pre-processing Approaches

Pre-processing strategies target disparities at the data stage, aiming to reduce inequities before model training begins. They mitigate bias within input data while preserving the predictive utility of relevant features. Foundational methods include reweighing, which adjusts instance weights to counteract imbalances across groups (Kamiran and Calders 2012), and the Disparate Impact Remover, which modifies features to weaken correlations with protected attributes while retaining predictive

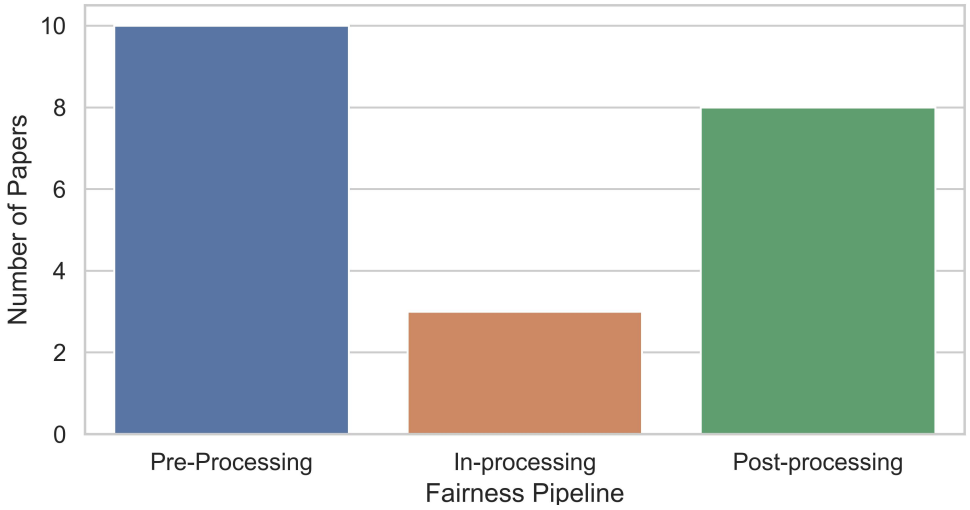

Figure 3.1: Overview mapping: Analysed papers and their focus on the three phases of the fairness pipeline.

validity (Feldman et al. 2015). Optimized pre-processing methods extend these approaches by explicitly transforming datasets to satisfy fairness constraints while bounding distortion and predictive loss (Calmon et al. 2017). Although these techniques are computationally efficient and model-agnostic, they risk attenuating informative signals and may undermine interpretability.

Resampling remains a central class of pre-processing interventions. Random oversampling duplicates minority-group instances, while SMOTE interpolates synthetic examples in feature space (Chawla et al. 2002). Empirical research shows that such techniques can improve statistical parity and equal opportunity but cautions that some distort subgroup distributions or smooth decision boundaries excessively, thereby amplifying bias rather than reducing it (Rančić, Radovanović, and Delibašić 2021; Zelaya 2019). To address these concerns, targeted methods such as FAWOS prioritize fairness-critical subgroups (Salazar et al. 2021), while Protected-Category SMOTE and ADASYN focus on underrepresented intersections (e.g., race $\times$ gender) to counter compounded disadvantage (Popoola and J. Sheppard 2024). Pipeline-level designs that embed these subgroup-sensitive samplers upstream of model training demonstrate measurable improvements at both aggregate and intersectional levels (Zhou, Kantarcioglu, and Clifton 2023; Kokhlikyan

et al. 2022). Beyond resampling, representation-based methods restructure the input space, training separate models for sub-populations (Jain, Huber, Fegaras, et al. 2019) or applying probabilistic transformations that balance discrimination control, distortion, and utility preservation (Pin Calmon et al. 2018). Collectively, these strategies underscore that mitigating bias at the data level requires explicit attention to the socio-technical and demographic contexts from which the data originate. Recent advances extend classical pre-processing by developing theoretically grounded and data-efficient frameworks that target equalized-odds fairness or privacy–fairness trade-offs. Yu et al. introduce FairBalance, a reweighting scheme that equalizes weighted class distributions across demographic groups to satisfy the necessary and sufficient conditions for relaxed equalized odds (Z. Yu, Chakraborty, and Menzies 2024). Unlike heuristic resampling methods, FairBalance derives its weighting function from formal conditions that minimize the smoothed Average Odds Difference on the training data, thereby offering provable fairness improvements with minimal utility loss across tabular and image datasets. Schrouff et al. advance this discussion by providing a causal-graphical analysis explaining why data balancing sometimes fails to achieve fairness or robustness. They demonstrate that balancing operations alter statistical dependencies without necessarily removing causal links between sensitive and outcome variables (Schrouff et al. 2024). Their results caution that pre-processing must align with the data-generating process; otherwise, balancing can introduce new biases or weaken model invariance.

A parallel research direction explores fair synthetic data generation as a means to achieve both bias mitigation and privacy preservation at the data level. Sikder et al. propose FAIR4FREE, a data-free distillation framework that learns fair latent representations via a variational autoencoder and then generates synthetic samples from noise without requiring access to real data (Sikder, Leng, and Heintz 2024). This approach produces high-fidelity synthetic datasets that satisfy demographic parity and equalized-odds criteria while maintaining predictive utility, addressing fairness and data-access limitations simultaneously. Complementing this, Liu et al. conduct

a comparative study of synthetic data generators and fairness algorithms within learning analytics, examining how models such as CTGAN, PATEGAN, ADSGAN, and DECAF mediate the triadic trade-off between privacy, fairness, and accuracy (Q. Liu et al. 2025). Their findings show that combining fairness-oriented pre-processing algorithms with synthetic data improves fairness more effectively than applying these algorithms to real data, underscoring the promise of synthetic generation as a fairness-enhancing pre-processing strategy. Collectively, these works broaden the scope of data-level fairness interventions by integrating causal reasoning, privacy-aware synthesis, and theoretically principled weighting within the pre-processing stage.

In the criminal justice domain, pre-processing has been the most extensively applied fairness phase for mitigating bias in recidivism risk prediction models. The literature reveals multiple approaches tailored to address entrenched racial and gender disparities inherent in criminal justice data. Jain et al. (Jain, Huber, Fegaras, et al. 2019) examined bias and accuracy in predicting prisoner recidivism using novel singular race models, which train and test sub-populations one at a time without explicitly including race as a feature. This design challenges standard practices that remove racial information to mitigate bias (Pin Calmon et al. 2018; Jain, Huber, Fegaras, et al. 2019). Compared with a base model trained on combined racial groups, the singular-race approach achieved higher accuracy but exacerbated false-positive disparities against African American defendants. The authors attribute this to dataset imbalance and limited feature diversity, suggesting that fairness cannot be achieved through racial partitioning alone. They propose that richer, behaviorally informed features, not just demographics, could help mitigate bias in future iterations.

Bhanu et al. (Jain, Huber, R. Elmasri, et al. 2020) introduced the Bias Parity (BP) Score, a fairness metric that quantifies model bias across different configurations of feature augmentation. Using temporal data from the "Recidivism of Prisoners Released in 1994" dataset, they showed that incorporating offenders' historical

trajectories improves both fairness and predictive accuracy, thereby challenging the conventional assumption that fairness necessarily trades off with accuracy. The BP Score enables comparison among models trained with varying history lengths, but incurs high computational cost due to repeated training and testing. Moreover, by relying on reconviction rather than rearrest as the outcome variable, the model risks reproducing biases inherent in judicial decision-making.

Similarly, Du et al. (Pin Calmon et al. 2018) proposed a probabilistic pre-processing framework to reduce discrimination by transforming data distributions while maintaining individual fairness and utility. Applied to the COMPAS dataset, this method adjusts protected attributes, non-protected attributes, and outcomes to satisfy group fairness, distortion, and utility constraints jointly. The results showed reduced disparate impact with some cost to accuracy, reaffirming that removing protected attributes alone is insufficient to prevent indirect discrimination (Pin Calmon et al. 2018; Jain, Huber, Fegaras, et al. 2019; A. Biswas and Mukherjee 2021). This approach demonstrates that fairness can be optimized probabilistically while preserving essential structural information in recidivism data. Expanding this line of research, Kobayashi et al. (Kobayashi and Nakao 2020) developed a One-vs-One intersectional mitigation strategy to address subgroup inequities ignored by traditional pre-processing methods. Using the COMPAS dataset, they divided gender and race into subgroups (e.g., female–nonwhite, male–white) and trained pairwise models to mitigate disparities between these intersections. Their findings indicate significant improvement in demographic parity, equalized odds, and equal opportunity without sacrificing accuracy, demonstrating that intersectionally aware pre-processing can outperform conventional single-axis techniques.

In a different direction, Dass et al. (Dass et al. 2022) used mugshots to generate or fill missing race information in criminal justice datasets, motivated by frequent data gaps and inaccuracies. They employed facial processing technology to infer race categories, thereby improving dataset completeness for fairness analysis. While this approach aids transparency in identifying bias sources, it raises ethical

concerns regarding biometric surveillance and privacy. Miron et al. (Miron et al. 2021) took a complementary approach by evaluating sources of bias within the Catalonia juvenile justice dataset, focusing on disparities affecting women and foreign nationals. They applied stratified oversampling and protected-feature removal as pre-processing strategies. They found that static features (e.g., demographics, past convictions) correlate more strongly with protected attributes than dynamic features (e.g., substance use, peer rejection). The study concluded that reliance on static features reproduces structural biases, whereas integrating dynamic, behavioral data promotes more equitable model outcomes.

Taken together, pre-processing interventions have advanced the technical capacity to mitigate bias in recidivism risk prediction, yet they remain limited in addressing structural inequities embedded in criminal justice data. While strategies such as subgroup-sensitive sampling, probabilistic transformation, and intersectional modeling have improved fairness across metrics, they often entail trade-offs between interpretability, accuracy, and ethical legitimacy. Data-level interventions cannot ensure justice-aware fairness in isolation because they operate downstream of historically biased institutional processes. Their effectiveness depends on how fairness constraints interact with domain-specific realities, underscoring the need for lifecycle-aware frameworks that integrate pre-processing with broader mechanisms of accountability and socio-technical oversight.

3.4.2 In-processing Approaches

In-processing methods have emerged as one of the most influential strategies for embedding fairness constraints directly into the model training process. Unlike pre-processing interventions that manipulate input data, in-processing techniques intervene at the algorithmic level to shape learning dynamics and enforce fairness-aware objectives. These approaches reflect a broader shift from reactive to proactive bias mitigation, seeking to internalize ethical considerations within model optimization itself (Wan et al. 2023).

A central line of research conceptualizes algorithmic fairness as an optimization problem balancing accuracy and equity. Corbett-Davies et al. (Corbett-Davies et al. 2017) formalized this tension through constrained and unconstrained optimization frameworks, demonstrating that achieving statistical parity often conflicts with maximizing public safety in recidivism prediction. Using the Broward County COMPAS dataset, they showed that constrained optimization, which applies race-specific thresholds to equalize outcomes, reduces racial disparities but may lead to the release of higher-risk defendants. Conversely, unconstrained models maintain safety at the cost of fairness. This seminal study revealed the ethical and practical dilemma at the heart of criminal justice algorithms, whether fairness should prioritize equality of outcomes or protection of the public. Their findings have informed ongoing policy debates about how AI models in high-stakes contexts should balance normative and utilitarian goals.

Building on this formulation, Bechavod and Ligett (Bechavod and Ligett 2017) introduced fairness regularization to minimize disparities in false positive and false negative rates across demographic subgroups. Their method incorporates penalty terms, based on absolute and squared differences, into logistic regression models, enabling direct control of error-rate parity during training. Tested on the COMPAS dataset, this approach achieved comparable predictive accuracy to baseline classifiers while delivering substantially improved fairness outcomes. The study also underscored the advantage of intervening during the in-processing phase rather than after training: embedding fairness constraints within the learning objective avoids the suboptimal adjustments often required in post hoc correction.

In parallel, Berk et al. (R. Berk, Heidari, Jabbari, M. Joseph, et al. 2017) extended fairness regularization to regression settings, introducing a convex framework that unifies group, intermediate, and individual fairness within a single optimization structure. Their model quantifies trade-offs between fairness and accuracy using the Price of Fairness (PoF) metric, a standardized measure for comparing fairness loss across datasets and regularizers. Applied to recidivism data, this framework

illustrated that fairness–accuracy trade-offs depend on both the chosen fairness definition and the underlying data distribution, highlighting the contextual sensitivity of fairness optimization in real-world decision systems. Collectively, these studies demonstrate that fairness cannot be defined in the abstract; it must be situated within the ethical and practical constraints of its application domain.

Beyond regularization, other research reconceptualizes in-processing fairness through algorithmic reformulations. Reductions-based approaches transform fairness-constrained learning into sequences of cost-sensitive classification tasks, allowing the enforcement of fairness criteria such as equalized odds across arbitrary model families (Agarwal et al. 2018). Adversarial debiasing extends this principle by training a predictor jointly with an adversary that attempts to infer protected attributes, thereby minimizing demographic leakage without sacrificing predictive performance (B. H. Zhang, Lemoine, and Mitchell 2018). Hybrid approaches that combine in-processing constraints with pre-processing interventions, such as subgroup-sensitive oversampling, demonstrate superior robustness by addressing both data imbalance and optimization bias simultaneously (Calmon et al. 2017; Ghosh, Genuit, and Reagan 2021). More recent extensions employ fairness-aware gradient adjustments to regulate model updates during training, offering promising, but still domain-limited, evidence of effectiveness in healthcare and social prediction contexts (X. Wang and C. C. Yang 2025).

Finocchiaro (Finocchiaro 2024) deepens the theoretical basis of in-processing fairness by employing the framework of property elicitation to analyze how fairness regularizers reshape a model's optimization landscape. The study establishes formal conditions under which regularized losses preserve or alter the elicited statistical property of an unregularized objective, thereby clarifying when fairness constraints fundamentally modify model behavior. Empirical tests on lending and health datasets confirm that increasing the strength of fairness regularization systematically shifts prediction boundaries and alters group outcomes, offering a principled explanation for the fairness–accuracy trade-offs observed in practice. Similarly, Ni et al. (Ni et

al. 2024) address fairness when sensitive attributes are unavailable by introducing a dual-model knowledge-sharing framework that regularizes bias through pseudo-learning between high- and low-confidence classifiers. Their method, evaluated on the COMPAS, New Adult, and CelebA datasets, demonstrates that fairness can be enhanced even in privacy-constrained settings, advancing fairness-aware optimization beyond explicit demographic supervision. Extending this direction, Zanna and Sano (Zanna and Sano 2024) combine multi-task learning, Monte-Carlo dropout, and Pareto optimality to jointly model uncertainty and fairness during training. Their approach automatically identifies Pareto-efficient solutions balancing accuracy and disparity, showing that fairness-aware optimization can be achieved through probabilistic calibration rather than explicit constraint formulation. Together, these studies expand in-processing fairness research toward theoretically grounded, data-efficient, and uncertainty-aware optimization strategies that integrate fairness objectives directly into model learning.

Taken together, the literature on in-processing fairness reflects a maturing understanding of how ethical principles can be encoded within algorithmic learning. The field has progressed from heuristic post hoc corrections toward formal optimization frameworks that operationalize fairness as a quantifiable design objective. Yet persistent challenges remain, including how to determine which fairness definition should guide optimization, how to assess the moral legitimacy of trade-offs between fairness and public safety, and how to ensure that mathematical parity corresponds to substantive justice. In high-stakes domains such as recidivism prediction, these tensions highlight the limits of purely technical solutions and underscore the need for governance structures that align algorithmic design with broader social accountability. The emerging consensus suggests that in-processing fairness must be embedded not only in code but also in the institutional practices that define what constitutes a “fair” outcome in the first place.

3.4.3 Post-processing Approaches

Post-processing approaches address fairness concerns at the final stage of model deployment by adjusting decision outputs after training to satisfy specified fairness constraints. These methods are particularly valuable in high-stakes domains such as criminal justice, where retraining models may be infeasible due to regulatory or institutional barriers. Conceptually, they represent pragmatic attempts to retrofit fairness into already trained systems, offering a degree of control over outcomes, though typically limited in their capacity to correct upstream representational or institutional biases.

Foundational contributions to post-processing include the Equalized Odds adjustment introduced by Hardt et al. (Hardt, Price, and Srebro 2016), which aligns false positive and false negative rates across demographic groups, and the Calibrated Equalized Odds method developed by Pleiss et al. (Pleiss et al. 2017), which balances calibration with error-rate parity under differing base rates. While these methods marked significant advances in practical fairness, they also revealed fundamental trade-offs: calibration and error-rate equality cannot both be satisfied when group-based rates differ. Subsequent refinements attempted to reconcile these conflicts. Biswas et al. (A. Biswas and Mukherjee 2021) proposed a combinatorial algorithm for Proportional Equality (CAPE) that accounts for prior probability shifts, where training and testing distributions diverge, through a quantification-based ensemble sampling method. Using the COMPAS dataset, CAPE outperformed several pre-, in-, and post-processing techniques, including Reweighing, Adversarial Debiasing, Meta-Fair algorithms, Calibrated Equalized Odds, and Reject Option Classification, demonstrating robustness across fairness metrics such as false positive/negative differences and overall predictive disparity. CAPE's focus on proportional equality highlights the need to account for distributional shifts, although its reliance on stable data poses challenges in dynamic environments where concept drift occurs.

Other scholars extend post-processing beyond statistical parity to consider sufficiency-based criteria such as predictive parity and group calibration, both widely used in

evaluating recidivism prediction instruments (Mishler, E. H. Kennedy, and Chouldechova 2021; X. Zeng, Dobriban, and G. Cheng 2022). Zeng et al. (X. Zeng, Dobriban, and G. Cheng 2022) developed FairBayes-DPP, an algorithm that operationalizes predictive parity through group-wise thresholding rules (GWTRs), establishing that all Bayes-optimal classifiers under predictive parity conform to this structure. In contrast, Mishler et al. (Mishler, E. H. Kennedy, and Chouldechova 2021) introduced counterfactual equalized odds, a criterion that allows users to negotiate explicit trade-offs between fairness and predictive accuracy. Using the COMPAS dataset, they showed that approximate counterfactual equalized odds more effectively mitigate disparate impact than predictive parity. However, their findings contradict Zeng et al.'s claim regarding the sufficiency of predictive parity. These debates illustrate ongoing theoretical tensions between observational and counterfactual notions of fairness and reveal that no single post-processing strategy can universally resolve the fairness–accuracy dilemma.

Efforts to unify multiple fairness definitions within a single framework have shaped the evolution of post-processing. Quadrianto et al. (Quadrianto and Sharmanska 2017) proposed a unified machine learning framework that integrates demographic parity, fairness through unawareness, equality of opportunity, and equalized odds by combining privileged learning, where certain features are available during training but not at test time, with distribution-matching techniques. This framework enables flexible alignment across fairness criteria. It offers a post hoc mechanism to balance trade-offs between them, an approach particularly useful for predictive systems such as recidivism risk assessment tools, where institutional constraints limit model retraining. Similarly, Zhao et al. (Zhao et al. 2019) introduced CFAIR, a conditional fair representation algorithm based on minimizing balanced error rate (BER) across protected characteristics and target variables. Evaluated on the COMPAS dataset, CFAIR achieved simultaneous equalized odds and accuracy parity by aligning conditional rather than marginal distributions, though its performance proved more stable on balanced datasets.

In practical recidivism prediction contexts, post-processing has also been used to evaluate the consequences of "blinding" models to sensitive attributes. Skeem et al. (J. Skeem and C. Lowenkamp 2020) demonstrated that excluding race or proxy features can paradoxically worsen calibration and inflate error disparities. Their study compared three debiasing algorithms, Proxy Elimination, Race Elimination, and Criminal History Discount, and found that incorporating structured criminal history information improves predictive validity while reducing error-rate imbalances. Jain et al. (Jain, Huber, R. A. Elmasri, et al. 2020) supported these findings, showing that using offender history enhances both accuracy and fairness when applied to the Recidivism of Prisoners Released in 1994 dataset. They developed a race-blind model that improves fairness through false-positive rate parity across sub-populations while maintaining predictive accuracy. Collectively, these studies suggest that omitting sensitive attributes, often presumed to ensure fairness, can instead obscure underlying disparities that calibrated post-processing approaches help reveal and manage.

Calibration itself remains a central focus of refinement. Pleiss et al. (Pleiss et al. 2017) emphasized the necessity of maintaining calibrated probability estimates while minimizing disparities in error rates. Their experiments using the COMPAS dataset demonstrated that risk instruments calibrated to racial subgroups can prevent judges from implicitly considering race in decision-making and reduce discrimination in false positive and false negative rates. However, they also confirmed that calibration is mathematically incompatible with equalized error rates. Karimi et al. (Karimi-Haghighi and Castillo 2021b) extended this work by applying calibrated Equalized Odds methods to the Catalan RisCanvi dataset using logistic regression, multilayer perceptron, and support vector machine models for violent and general recidivism. Calibration across nationality and age subgroups reduced false-positive disparities, highlighting the potential of domain-specific calibration strategies for improving fairness. These results reinforce that calibration, though constrained by theoretical limits, remains an essential mechanism for fairness alignment when adapted to local

data and institutional contexts.

Recent research further broadens post-processing fairness by incorporating optimality, continuity, and stability principles. Zhao (Zhao 2024) formalizes post-processing as a theoretically optimal framework grounded in optimal transport, demonstrating that Bayes-consistent post-processing can achieve fairness–accuracy trade-offs equivalent to in-processing approaches across binary, multiclass, and multi-group tasks. Small et al. (Small et al. 2024) advance this direction by introducing equalized individual odds, a continuous Lipschitz-constrained post-processing method that reconciles group and individual fairness through smooth probabilistic mapping. Di Gennaro et al. (Di Gennaro et al. 2024) contribute Ratio-Based Model Debiasing (RBMD). This model-agnostic technique rescales prediction logits to satisfy fairness constraints while minimizing deviation from original outputs, thereby enhancing interpretability and stability. Pinkava et al. (Pinkava, McFarland, and Mashhadi 2024) propose a reinforcement-learning-based post-processing system that iteratively reduces equalized-odds violations in black-box models using synthetic queries, establishing fairness-as-a-service without retraining or data access. Extending these ideas, Eberhard et al. (Eberhard et al. 2025) develop a Fairness Adjuster framework that learns correction functions on model predictions to replicate adversarial debiasing outcomes post hoc, demonstrating near-equivalent fairness–accuracy performance to full retraining. Complementing these algorithmic advances, the Efficient Post-Processing for Equal Opportunity framework (Xian and Zhao n.d.) extends equal opportunity to multiclass settings via linear programming, offering provable guarantees on fairness–utility trade-offs and computational scalability. Together, these studies reposition post-processing from a reactive correction to a theoretically grounded and computationally rigorous fairness mechanism capable of bridging model deployment and ethical accountability.

Collectively, these post-processing studies illustrate both the adaptability and the constraints of fairness adjustments made after model training. Algorithms such as CAPE, FairBayes-DPP, CFAIR, and unified post-hoc frameworks demonstrate

that fairness can be introduced without retraining. Yet, their success often depends on stable distributions and clear subgroup definitions, conditions rarely guaranteed in dynamic, high-stakes environments such as criminal justice. Moreover, cross-jurisdictional analyses (Karimi-Haghighi and Castillo 2021b) show that techniques effective in one context may fail in another, emphasizing the importance of local validation and contextual awareness. Post-processing thus remains an essential yet partial tool: it can mitigate observable disparities and support regulatory compliance, but it cannot resolve the upstream causes of unfairness rooted in data generation, institutional bias, or historical inequities. Addressing these deeper issues requires integrated, lifecycle-wide approaches that combine pre-, in-, and post-processing interventions within justice-aware frameworks for algorithmic fairness.

3.4.4 Toward Integrated Pipelines

Across these strands of research, a consistent finding emerges: single-stage fairness interventions tend to be brittle. Comparative studies show that no single method consistently outperforms others across datasets or fairness criteria, underscoring the need for integrated, pipeline-level approaches (Friedler et al. 2019; Caton and Haas 2024). More recent work suggests that multi-stage strategies yield more stable trade-offs between fairness and accuracy, particularly under conditions of subgroup imbalance or varying base rates (Z. Chen et al. 2023; Caton and Haas 2024). Rather than relying on isolated fixes, multi-stage pipelines combine subgroup-sensitive sampling, fairness-aware training, and post-hoc adjustments, aligning interventions across the learning process. By coordinating methods in this way, such pipelines improve performance across multiple metrics and reveal intersectional disparities that remain hidden in aggregate evaluations, echoing concerns raised in applied studies of recidivism and healthcare (Jain, Huber, R. Elmasri, et al. 2020; J. Skeem and C. Lowenkamp 2020; Karimi-Haghighi and Castillo 2021b).

In summary, bias mitigation strategies span every stage of the machine learning pipeline, from data preparation to model training and output adjustment. Each

class of methods addresses distinct sources of bias, yet none offers a complete solution in isolation. The growing consensus favors integrated pipelines that combine interventions, embed intersectional audits, and account for domain-specific structures. This integrated view provides a conceptual bridge to the next section, which explores intersectionality and subgroup fairness as essential to understanding how compounded harms persist even when conventional fairness metrics are satisfied.

3.5 Intersectionality and Subgroup Fairness

Traditional approaches to fairness focus on single-axis protected attributes such as race or gender. This perspective risks overlooking the compounded disadvantages experienced by individuals situated at the intersections of multiple marginalized identities. Intersectionality, first articulated by Crenshaw (Crenshaw 2022), highlights how overlapping systems of oppression, racism, sexism, and classism, produce harms that single-attribute analyses cannot adequately capture. In machine learning, this means that models appearing fair when evaluated on isolated attributes may nonetheless disproportionately disadvantage intersectional subgroups (L. T. Liu et al. 2018).

The risk of such concealed harms is evident in the phenomenon of fairness gerrymandering. Kearns et al. (Kearns et al. 2018) demonstrated that algorithms can satisfy fairness constraints across individual attributes while systematically disadvantaging subgroups defined by their intersections, for example, penalizing Black women even when overall gender and race parity appears satisfied. Their subgroup auditing framework introduced tools to evaluate fairness across a wide range of subgroup combinations, motivating more granular evaluation. Building on this work, Ghosh et al. (Ghosh, Genuit, and Reagan 2021) proposed worst-case subgroup fairness, which explicitly optimizes for the most disadvantaged groups. In recidivism prediction, Jain et al. (Jain, Huber, Fegaras, et al. 2019; Jain, Huber, R. Elmasri, et al. 2020) show how subgroup-sensitive audits and feature-rich modeling reveal disparities masked by aggregate evaluations. Recent research in healthcare and other

high-stakes domains similarly demonstrates that interventions appearing successful at the aggregate level can still harm intersectional minorities (Buolamwini and Gebru 2018; Mehrabi et al. 2021; Hanna et al. 2020a).

Complementary approaches extend fairness guarantees across complex subgroup structures. Foulds et al. (Foulds et al. 2020) introduced differential fairness, a statistical measure that ensures deviations in outcomes across intersectional subgroups remain bounded. Similarly, Hebert-Johnson et al. (Hébert-Johnson et al. 2018) developed multicalibration, which enforces calibration across computationally identifiable subgroups. These techniques move beyond group averages to offer guarantees across exponentially many subgroup combinations. Yet their application remains constrained: they are computationally intensive, require large and diverse datasets, and can still reproduce inequities when underlying data are biased. Scholars therefore argue that technical guarantees must be complemented by systemic or pipeline-level interventions, aligning fairness auditing with data preprocessing, model design, and governance practices (Barocas, Hardt, and Narayanan 2023; Mehrabi et al. 2021; Mitchell et al. 2019).

Empirical studies across high-stakes domains highlight the salience of intersectionality. In computer vision, Buolamwini and Gebru (Buolamwini and Gebru 2018) revealed that commercial facial recognition systems performed worst on Black women, catalyzing awareness of compounded harms. In criminal justice, Kobayashi et al. (Kobayashi and Nakao 2020) developed a one-vs.-one strategy to mitigate subgroup disparities in COMPAS, while Miron et al. (Miron et al. 2021) documented disproportionate harms to nationality–gender subgroups in juvenile justice predictions. Dass et al. (Dass et al. 2022) warn against relying on proxy variables such as mugshots for subgroup identification, noting that such proxies may fill demographic gaps but risk reinforcing stereotypes. Similar findings emerge in healthcare: Valentine et al. (Valentine, Charney, and Landi 2024) show that diagnostic fairness depends on accounting for intersections of race, sex, and socioeconomic status, while Ramachandranpillai et al. (Ramachandranpillai et al. 2024) find that multimodal

prediction models in intensive care disproportionately harm subgroups when clinical notes and imaging data are unevenly distributed. Research in financial services similarly reveals that auditing for subgroup intersections uncovers disparities invisible under single-axis evaluations (S. Kim et al. 2023). Extending these insights, recent scholarship demonstrates that systematically embedding intersectional subgroup auditing into fairness pipelines improves outcomes for disadvantaged populations often neglected in conventional evaluations (Hanna et al. 2020a; Mehrabi et al. 2021; Foulds et al. 2020).

Alongside methodological advances, normative critiques insist that intersectional fairness cannot be reduced to technical metrics alone. Binns (Binns 2018) argues that most fairness interventions neglect structural injustice, contending that prioritizing the worst-off better aligns with egalitarian principles. Birhane (Kasy and Abebe 2021) critiques fairness research for overlooking entrenched power imbalances and calls for the inclusion of historically marginalized voices in AI governance. Hanna et al. (Hanna et al. 2020b) advocate race-conscious fairness frameworks grounded in critical race theory. Complementing these perspectives, scholars highlight justice-aware approaches that integrate technical interventions with ethical reflection, participatory governance, and gender-conscious analysis (Jobin, Ienca, and Vayena 2019; Floridi 2019; Thiebes, Lins, and Sunyaev 2021). Related work in recidivism prediction emphasizes that fairness cannot be separated from broader requirements of trustworthy AI, including transparency, accountability, and explainability (Lo Piano 2020; McKay 2020; Ződi 2022). Taken together, these contributions reinforce that intersectional fairness is fundamentally a socio-technical challenge.

The evidence demonstrates that intersectionality reshapes both the methodological and normative landscape of fairness research. While existing tools such as subgroup auditing, differential fairness, and multicalibration represent important advances, they remain constrained by computational demands, data limitations, and unresolved ethical debates. Addressing these challenges requires approaches

that embed intersectional auditing into fairness pipelines and situate technical innovations within broader socio-technical frameworks. The next section examines integrated fairness frameworks that combine technical, ethical, and regulatory strategies into unified approaches for achieving equitable AI in practice.

3.6 Integrated Fairness Frameworks

The limitations of isolated interventions have motivated a growing interest in integrated fairness frameworks that combine methods across multiple stages of the machine learning pipeline. These frameworks aim to reconcile competing fairness metrics, preserve predictive performance, and mitigate subgroup disparities in real-world, high-stakes contexts. While promising, many remain validated only on benchmark datasets and devote limited attention to the socio-technical dimensions of fairness (Friedler et al. 2019; Bellamy et al. 2019).

At the pipeline level, Chen et al. (Z. Chen et al. 2023) conducted one of the most extensive evaluations to date, benchmarking 17 mitigation strategies across diverse decision-making tasks. Their findings indicate that single-stage interventions rarely achieve robust fairness, whereas multi-stage combinations, such as pre-processing paired with in-processing constraints, yield more stable outcomes. Similarly, Caton et al. (Caton and Haas 2024) advocate fairness-by-design approaches that embed mitigation throughout the pipeline. Extending this trajectory, scholars emphasize integrated designs that coordinate pre-processing, in-processing, and post-processing methods to achieve more reliable trade-offs between fairness and accuracy, particularly when evaluated across intersectional subgroups (Kamiran and Calders 2012; Feldman et al. 2015; Zafar et al. 2017). In criminal justice specifically, Eckhouse et al. (Eckhouse et al. 2019) argue that bias must be addressed simultaneously at the data, model, and evaluation stages, reinforcing the logic of integrated design.

Beyond heuristic combinations, several frameworks adopt causality-driven and adaptive strategies. Zhang et al. (M. Zhang and Sun 2022) introduce adaptive fairness regularization grounded in causal inference, arguing that conventional ap-

proaches overlook structural relationships between features and outcomes. Quadrianto et al. (Quadrianto and Sharmanska 2017) offer a unified model that reconciles multiple fairness definitions by recycling privileged information through distribution matching. Related work emphasizes counterfactual reasoning as a mechanism for embedding causal assumptions directly into fairness constraints (Kusner et al. 2017; Kilbertus et al. 2020). Although theoretically compelling, such methods remain computationally intensive and underexplored in high-stakes domains such as criminal justice. More recent extensions, including Wang and Yang's FairGrad (X. Wang and C. C. Yang 2025), adapt these ideas to healthcare by aligning gradient updates with subgroup fairness objectives in sepsis prediction, though applications remain narrowly scoped.

Integrated fairness frameworks also increasingly employ multi-objective optimization. Kozodoi et al. (Kozodoi, Jacob, and Lessmann 2022) apply Pareto optimization in credit scoring, generating non-dominated solutions that balance fairness, accuracy, and profitability. Chakraborty et al. (Chakraborty et al. 2020) introduce the Fairway framework, which combines pre- and in-processing through multi-objective optimization. Building on these approaches, recent work in criminal justice prediction demonstrates that fairness constraints such as demographic parity, equalized odds, and error-rate balance can be jointly optimized to improve subgroup equity with limited costs to accuracy (R. Berk 2019; Menon and Williamson 2018). Related studies by Martinez et al. (G. Yu et al. 2025) and Li et al. (Nagpal et al. 2024) show that Pareto-efficient solutions make fairness–accuracy trade-offs more transparent. Still, even optimized pipelines can perpetuate subgroup harms, underscoring the importance of embedding intersectional auditing within optimization frameworks (Ghosh, Genuit, and Reagan 2021).

Domain-specific adaptations further illustrate the potential of integrated pipelines. In healthcare, Raza et al. (Raza, Pour, and Bashir 2023) apply multi-stage strategies to diabetic retinopathy prediction, finding that neither pre- nor in-processing alone suffices to address systemic disparities. Ramachandranpillai et al. (Ramachandran-

pillai et al. 2024) extend this work by evaluating fairness across multimodal ICU tasks in MIMIC-IV, highlighting subgroup-specific harms across demographic intersections and data modalities. In language technologies, Wang et al. (Chuqiao Wang, Li, and R. Zhang 2024) propose FairLM, an integrated framework that combines fairness-aware active learning with structural constraints to mitigate representational harms in large language models. Within criminal justice, systematic reviews observe that existing fairness pipelines rarely integrate other pillars of trustworthy AI such as transparency, accountability, and explainability (Lo Piano 2020; McKay 2020; Ződi 2022). Scholars therefore emphasize the need to involve practitioners and stakeholders in pipeline design, arguing that their exclusion undermines accountability and long-term legitimacy (Cadigan and C. T. Lowenkamp 2011; Delgado et al. 2023).

Taken together, these literatures suggest that integrated fairness frameworks represent one of the most promising paths forward for achieving equitable AI in high-stakes domains. By combining interventions across stages, incorporating causal reasoning, leveraging multi-objective optimization, and tailoring solutions to domain contexts, these frameworks move closer to addressing both the technical and intersectional dimensions of fairness. Yet they also highlight persistent gaps, particularly in external validity, intersectional generalization, and socio-technical embedding, that this thesis seeks to address. The next chapter builds on these insights by developing multi-stage, intersectionally aware, and socio-technically grounded fairness pipelines, advancing both the empirical and normative foundations for trustworthy AI.

3.7 Research Gaps and Positioning of This Thesis

The review of trustworthy AI, algorithmic fairness, bias mitigation strategies, intersectionality, and integrated frameworks reveals five critical gaps that remain unresolved in the current literature. These gaps carry particular significance in high-stakes domains such as criminal justice, finance, and healthcare, where fairness

intersects with broader socio-technical concerns of transparency, accountability, and legitimacy. Addressing these gaps is essential for developing fairness-aware systems that are not only technically robust but also ethically and institutionally credible.

The following subsections identify and synthesize these five gaps, emphasizing how they constrain the practical and normative advancement of fairness research and motivating the methodological contributions developed in the next chapter.

Gap 1: Operationalizing Trustworthy-AI Principles Through Consistency, Reliability, Explainability, and Interpretability. Existing frameworks for trustworthy AI, including the EU High-Level Expert Group (HLEG) Guidelines (H. AI 2019), the NIST AI Risk Management Framework (National Institute of Standards and Technology 2023), and the OECD AI Principles (OECD 2019), serve as influential global references that define broad values of fairness, accountability, transparency, and robustness. Despite their prominence, numerous analyses show that these frameworks often remain aspirational, offering limited guidance on how to operationalize ethical principles across the AI lifecycle (Jobin, Ienca, and Vayena 2019; Fjeld et al. 2020; Brundage et al. 2020). This gap between principle and practice restricts the ability of developers, auditors, and regulators to assess whether AI systems demonstrate genuine trustworthiness in real-world operation.

A recurring limitation across these frameworks is the lack of clear operational mechanisms for translating key requirements, *reliability*, *explainability*, *interpretability*, and *consistency*, into measurable and auditable practices. Although the literature recognizes these dimensions as foundational to trustworthy AI (Doshi-Velez and B. Kim 2017; Gilpin et al. 2018; Miller 2019; Varshney 2016), researchers and institutions often address them in isolation rather than as interdependent elements of a coherent governance process. When treated separately, efforts to promote reliability, transparency, or interpretability become fragmented across lifecycle stages, providing little assurance that declared ethical or regulatory commitments persist in practice. This fragmentation erodes both public confidence and institutional accountability, particularly in high-stakes domains such as healthcare, finance, and

criminal justice (Raji and Buolamwini 2019; Veale, Van Kleek, and Binns 2018; Suresh and Guttag 2019).

As discussed in Chapter 2, reliability, explainability, and interpretability have been widely defined, yet their interdependencies and lifecycle alignment remain under-theorized. Building on these foundations, **this thesis proposes an Extension of the Trustworthy AI Framework** that integrates *consistency, reliability, explainability, and interpretability* as interdependent, lifecycle-oriented requirements. The proposed framework (Fig. 3.2) positions these four requirements not as isolated technical criteria but as mutually reinforcing conditions that must remain aligned throughout system development, deployment, and monitoring. Consistency functions as the connective principle that maintains this alignment across the AI lifecycle.

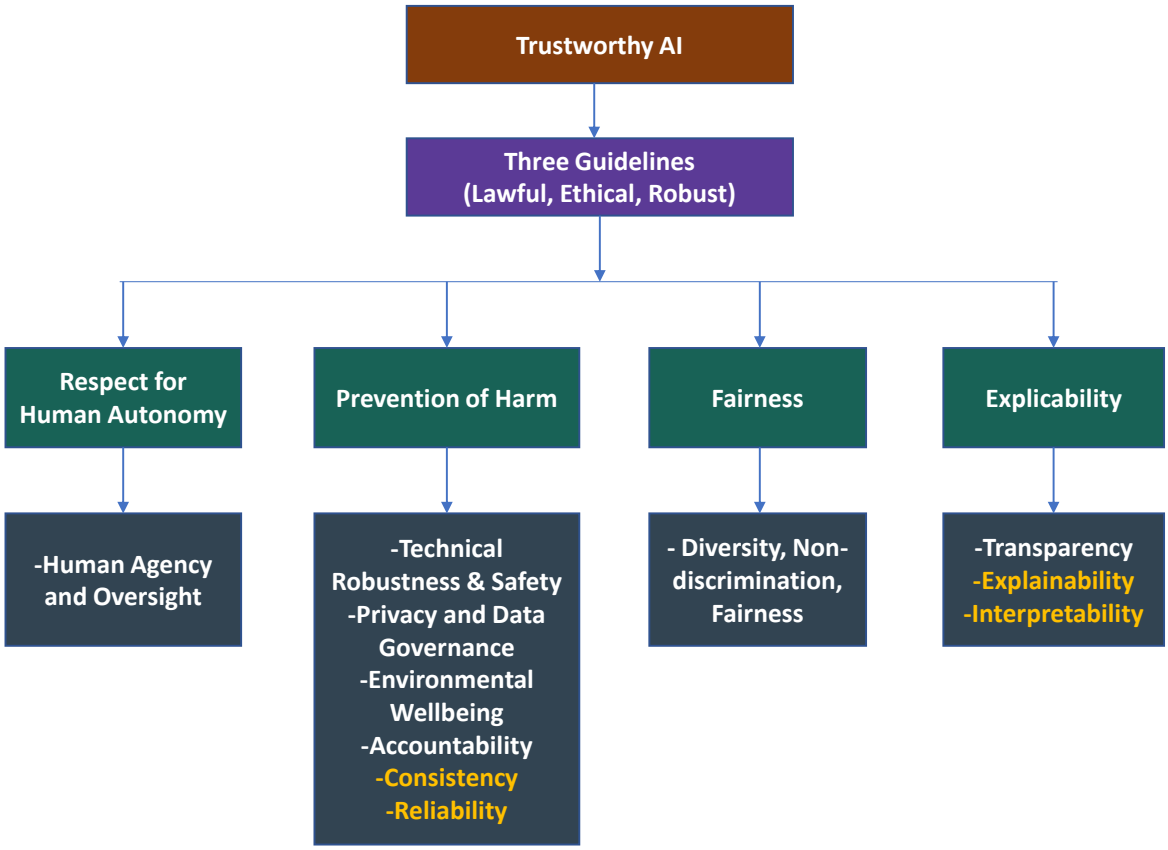

Figure 3.2: Extension of the Trustworthy AI Framework proposed in this thesis, integrating *Consistency*, *Reliability*, *Explainability*, and *Interpretability* as interdependent requirements across the AI lifecycle.

By embedding these four requirements within a unified and auditable structure, the proposed framework reconceptualizes trustworthy AI as an evolving system of commitments that sustain ethical and technical integrity over time. It establishes evaluative criteria that assess whether technical performance, transparency mechanisms, and governance processes remain coherent and traceable under real-world conditions. Hence, it strengthens and lays the conceptual foundation for the fairness- and justice-aware analyses developed in the subsequent gaps and empirical chapters.

Gap 2: Fragmentation of fairness interventions across the pipeline. Most studies implement fairness-enhancing techniques at a single stage of the machine learning lifecycle. Pre-processing approaches such as reweighing (Kamiran and Calders 2012) and the Disparate Impact Remover (Feldman et al. 2015) adjust data but offer no guarantees during training or prediction. In-processing methods such as adversarial debiasing (B. H. Zhang, Lemoine, and Mitchell 2018) optimize fairness during training but remain brittle under distributional shifts. Post-processing adjustments such as equalized odds (Hardt, Price, and Srebro 2016) modify model outputs but cannot address upstream bias. Recent research on integrated pipelines (Z. Chen et al. 2023; Caton and Haas 2024) demonstrates that multi-stage interventions achieve more stable fairness–accuracy trade-offs, yet such frameworks remain uncommon. This fragmentation highlights the need for robust, end-to-end fairness pipelines, a challenge that this thesis addresses in Chapter 5.

Gap 3: Insufficient attention to intersectionality and subgroup fairness. Although formalizations such as fairness gerrymandering (Kearns et al. 2018), worst-case subgroup fairness (Ghosh, Genuit, and Reagan 2021), and differential fairness (Foulds et al. 2020) provide theoretical foundations, most empirical studies evaluate fairness only at single-attribute or aggregate levels. Evidence from criminal justice (Kobayashi and Nakao 2020; Miron et al. 2021), healthcare (Valentine, Charney, and Landi 2024; Ramachandranpillai et al. 2024), and finance (Kozodoi, Jacob, and Lessmann 2022) shows that intersectional minorities continue to experience dispro-

portionate harm even when aggregate fairness improves. While recent scholarship emphasizes the importance of embedding subgroup and intersectional audits within fairness pipelines, systematic adoption of these approaches remains rare, leaving vulnerable populations underprotected (Buolamwini and Gebru 2018; Hanna et al. 2020a; Mehrabi et al. 2021). This gap is addressed in Chapter 7.

Gap 4: Data imbalance and fairness-aware augmentation. A persistent source of unfairness arises from imbalances in class labels and subgroup representation. Many datasets in criminal justice and healthcare exhibit skewed distributions; for example, minority groups are underrepresented in training data or disadvantaged by oversimplified sampling (Binns 2020). Although resampling and synthetic data generation methods such as SMOTE (Chawla et al. 2002) and FAWOS (Salazar et al. 2021) aim to mitigate these issues, their adoption remains fragmented, often confined to single datasets and lacking systematic evaluation of intersectional effects. Moreover, naïve augmentation can amplify bias when it fails to account for subgroup structures or reinforces spurious correlations. This gap underscores the need for fairness-aware augmentation strategies that explicitly address class imbalance, subgroup diversity, and domain-specific constraints, a challenge that this thesis examines in Chapter 6.

Gap 5: Reliance on single fairness metrics and calibration as a benchmark. Calibration often serves as the gold standard for fairness in high-stakes domains (Pleiss et al. 2017; Machado et al. 2024). However, calibrated models can still reproduce structural inequities when outcomes reflect biased social processes such as over-policing or discriminatory health cost proxies (Eubanks 2018; Obermeyer et al. 2019). Researchers demonstrate that calibration alone fails to prevent gendered and intersectional harms (Buolamwini and Gebru 2018; Hu 2025). Although extensions such as multicalibration (Hébert-Johnson et al. 2018) provide broader guarantees, purely technical remedies remain limited unless fairness is embedded within a sociotechnical and justice-aware design philosophy. More broadly, much of the fairness

literature evaluates interventions against a single metric at a time, overlooking how improvements in one dimension can exacerbate disparities in others. This narrow focus risks producing interventions that satisfy isolated criteria while leaving systemic or intersectional inequities unresolved. The persistent reliance on isolated metrics such as calibration underscores the need to re-envision fairness as a socio-technical and justice-oriented practice. The next subsection addresses this issue by situating calibration within a broader framework for Justice-Aware Fair ML design.

Beyond Calibration: Toward Justice-Aware Fair ML Design

Among the many fairness criteria proposed in the literature, calibration has attracted particular attention. Calibration ensures that predicted probabilities correspond to observed outcomes, for example, individuals assigned a 30% recidivism risk should, on average, reoffend at that rate. This property is appealing because it creates an interpretable link between model predictions and real-world outcomes, making it especially valuable in high-stakes domains such as criminal justice and healthcare (Pleiss et al. 2017; Kleinberg, Mullainathan, and Raghavan 2016).

Calibration often serves as a practical marker of both reliability and fairness: it enables decision-makers to interpret risk scores consistently across groups and remains relatively straightforward to measure. These characteristics explain its prominence in policy discussions and its adoption in widely deployed risk assessment systems.

Yet calibration's strengths also constitute its limitations. In contexts where historical data reflect systemic inequities, such as racially skewed policing or gendered disparities in healthcare, calibration can faithfully reproduce those inequities rather than mitigate them (Angwin et al. 2019; Obermeyer et al. 2019). A system may achieve perfect calibration yet still generate discriminatory outcomes if the underlying data encode biased institutional practices. This paradox demonstrates that calibration, while valuable, cannot serve as the sole or definitive indicator of fairness in trustworthy AI.

Moreover, calibration often conflicts with other fairness criteria. The impossibility results of Kleinberg et al. (Kleinberg, Mullainathan, and Raghavan 2016) and Chouldechova (Chouldechova 2017) show that when base rates differ across groups, it is mathematically impossible to simultaneously satisfy calibration and equalized error rates (e.g., equal false positives and false negatives). In practice, this means a recidivism prediction tool may remain well calibrated yet still disproportionately recommend detention for one group over another, as observed in the case of the COMPAS system (Angwin et al. 2016b). These trade-offs underscore that calibration, while useful, cannot alone guarantee fairness.

Building on this critique, this thesis advances a justice-aware framework that re-situates calibration within a broader socio-technical conception of fairness. To address these limitations, it introduces the **Justice-Aware AI Fairness (JAAF)** framework (Fig. 3.3). JAAF acknowledges the interpretability benefits of calibration but embeds them within a justice-centered approach. The framework integrates technical fairness interventions with intersectional analysis, participatory governance, and socio-technical accountability. Rather than rejecting calibration, JAAF repositions it as one component within a wider ecosystem of fairness practices. It rests on six mutually reinforcing pillars:

1. **Recognizing the limits of metrics**: Fairness metrics, including calibration, demographic parity, and equalized odds, encode normative assumptions and require critical application within their historical and institutional contexts. Practitioners must interrogate what these metrics measure, whose values they reflect, and which forms of injustice they may obscure.

2. **Center intersectionality**: Fairness must account for overlapping social identities such as race, gender, class, and disability. Without intersectional auditing, systems risk reproducing "fairness gerrymandering" effects that disproportionately harm marginalized subgroups (Kearns et al. 2018).

3. **Integrate fairness-enhancing interventions**: Pre-, in-, and post-processing

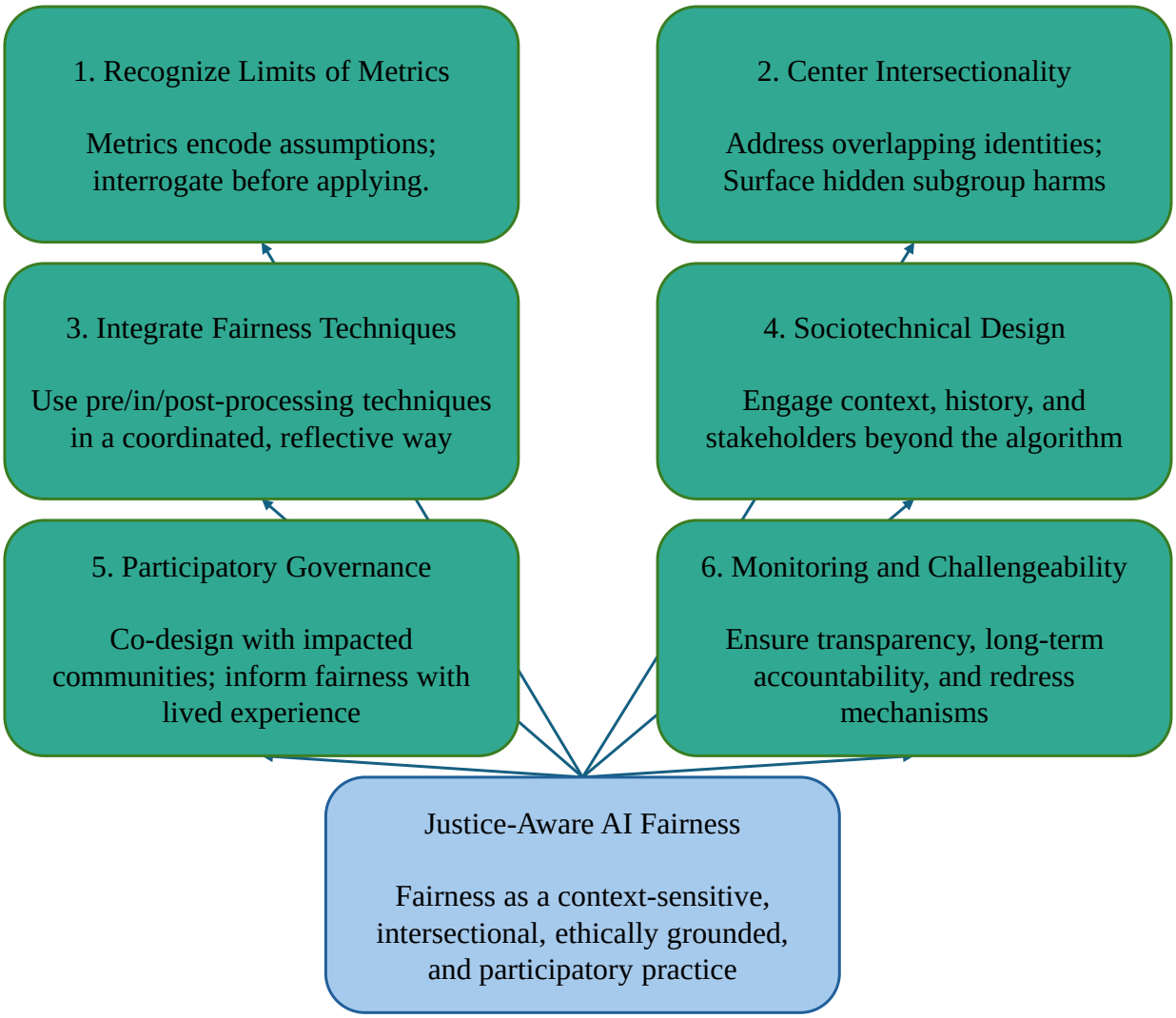

Figure 3.3: Expanded Framework for Justice-Aware AI Fairness (JAAF).

methods, such as reweighing, adversarial debiasing, and threshold adjustment, should operate jointly across the machine learning pipeline to *deliver* consistent and verifiable fairness guarantees (Caton and Haas 2023).

4. **Embed socio-technical design**: Fairness extends beyond algorithmic outputs; it emerges from data collection practices, institutional logics, and underlying power relations (Selbst et al. 2019). Designing fair systems therefore requires aligning technical objectives with broader social and organizational structures.

5. **Ensure participatory governance**: Impacted communities must participate in the design, deployment, and auditing of AI systems. Their lived

experiences offer contextual insights into harms and inequities that technical metrics alone cannot reveal (Costanza-Chock 2020).

6. **Commit to monitoring and challengeability**: Fairness demands ongoing audits, transparent documentation (e.g., model cards, datasheets), and mechanisms that enable contestation and redress when harms occur (Raji and Buolamwini 2019).

For instance, a recidivism prediction tool may appear calibrated across racial groups yet assign disproportionately higher risk scores to Black women because of intersectional data imbalances. Within the JAAF framework, such harms are identified through subgroup auditing, addressed through participatory consultation with affected communities, and documented using structured fairness reporting tools. A similar pattern emerges in healthcare, where calibrated models trained predominantly on male patient data systematically underestimate risks for women and non-binary patients (Obermeyer et al. 2019).

By reframing fairness as a justice-oriented, participatory, and intersectional commitment, JAAF challenges the overreliance on calibration and other isolated metrics. It embeds fairness across the AI lifecycle so that systems do not merely replicate historical inequities but actively foster more just and inclusive futures. The JAAF framework grounds the methodological design and case analyses presented in Chapters 5 and 7.

Gap 6: Weak external validity and limited cross-domain generalization. Most fairness pipelines undergo evaluation on narrow benchmark datasets, particularly COMPAS in the criminal justice domain. Although these datasets have driven significant methodological progress, they offer limited representation of the diversity of jurisdictions, institutional practices, and population structures encountered in real-world settings. Consequently, models and mitigation strategies that perform well on benchmark data often fail to generalize when deployed in new contexts, where base rates, data quality, and governance structures differ substantially.

Moreover, fairness interventions rarely incorporate domain-specific biases. Health data, for instance, embed inequities through cost-based proxies (Obermeyer et al. 2019), whereas criminal justice data reflect systemic over-policing of minority communities (Eubanks 2018). Methods validated in one domain often exhibit inconsistent behavior when transferred to another, demonstrating that fairness is not domain-neutral but deeply shaped by the social processes underpinning data generation.

Finally, external validity depends on socio-technical robustness. Scholars warn that evaluations focused narrowly on fairness–accuracy trade-offs neglect the broader governance dimensions that determine whether models earn public trust and remain legitimate in practice (E. P. Goodman and Trehu 2022). Bridging this gap requires fairness pipelines that undergo stress testing across diverse datasets and domains and operate within institutional and regulatory contexts. Only by doing so can gains in technical fairness translate into substantively equitable outcomes in the real world.

Positioning of this thesis. This thesis directly responds to the six research gaps identified above through four interrelated contributions. First, it addresses the lack of clear operational mechanisms for reliability, explainability, interpretability, and consistency by extending the EU Trustworthy AI framework (Gap 1). This extension embeds these dimensions as interdependent, lifecycle-oriented requirements, thereby transforming them from abstract ethical principles into auditable and actionable system properties.

Second, to mitigate the fragmentation of fairness interventions across the machine learning pipeline (Gap 2), the thesis develops and evaluates integrated fairness pipelines that combine pre-, in-, and post-processing interventions within multi-objective optimization frameworks. These pipelines overcome the brittleness of single-stage methods by chaining interventions across multiple points of the ML lifecycle, thereby enabling more stable and context-sensitive trade-offs between fairness and accuracy.

Third, in response to the insufficient attention to intersectionality and subgroup fairness (Gap 3), the thesis introduces systematic intersectional subgroup auditing. Rather than evaluating models solely on aggregate or single-axis metrics, the proposed approach explicitly captures compounded disadvantages among overlapping identities such as race $\times$ gender $\times$ age. This inclusion ensures that fairness interventions protect the most vulnerable subgroups, advancing both methodological rigor and social relevance.

Fourth, the thesis addresses data imbalance and fairness-aware augmentation (Gap 4) by investigating oversampling and augmentation strategies tailored to high-stakes domains. These strategies preserve subgroup structure to ensure that synthetic data generation improves fairness without reinforcing spurious correlations or exacerbating bias. This contribution demonstrates that fairness must be integrated into data generation and augmentation processes rather than treated as a post hoc correction.

To counter the persistent reliance on single fairness metrics (Gap 5), the thesis adopts a multi-metric evaluation framework that recognizes fairness as a plural and socio-technical construct. This framework acknowledges that improvements in one metric may diminish others and situates evaluation within a justice-aware design philosophy emphasizing participatory governance, monitoring, and challengeability.

Finally, to overcome weak external validity and limited cross-domain generalization (Gap 6), the thesis applies the proposed frameworks across diverse datasets in criminal justice (COMPAS, RisCanvi), finance (Adult Income), and healthcare (Irish Health Insurance). By evaluating fairness pipelines in settings that differ in population structures, institutional practices, and governance contexts, the research demonstrates both technical robustness and socio-technical relevance.

Taken together, these contributions move the field beyond fragmented and narrowly technical fixes, establishing an empirically validated, intersectionally aware, and socio-technically grounded framework for equitable and trustworthy AI.

3.8 Summary

This chapter reviews the literature on trustworthy AI, algorithmic fairness, bias mitigation strategies, intersectionality and subgroup fairness, and integrated fairness frameworks, with particular attention to high-stakes domains such as criminal justice, healthcare, and finance. Several overarching insights emerge from this review.

First, fairness definitions remain diverse, often mathematically incompatible, and deeply context-dependent. Fairness cannot be reduced to a single universal metric; rather, it must be understood as a pluralistic and socio-technical construct. Second, mitigation strategies operate at different stages of the machine learning pipeline, pre-processing, in-processing, and post-processing, yet they are often applied in isolation, producing fragmented and partial solutions. Third, despite advances in subgroup and intersectional fairness, most evaluations focus on single-axis attributes, overlooking compounded disadvantages that affect intersectional minorities. Fourth, the literature continues to rely heavily on narrow benchmark datasets, limiting external validity and raising concerns about generalizability across domains and populations. Finally, technical interventions tend to emphasize fairness–accuracy trade-offs while neglecting broader dimensions of trustworthy AI such as accountability, explainability, and institutional context.

Taken together, Sections 3.2 through 3.6 trace a progression from abstract ethical principles to technical interventions and integrated frameworks while revealing persistent conceptual and empirical challenges. The next section (3.7) synthesizes these challenges into six research gaps that frame the core contributions of this thesis, establishing the foundation for the methodological agenda developed in the chapters that follow.

Publication(s) Arising from this Chapter

The work presented in this chapter has been published in the following outlets:

1. Michael Mayowa Farayola, Irina Tal, Regina Connolly, et al. (2023). "Ethics and Trustworthiness of AI for Predicting the Risk of Recidivism: A Systematic Literature Review". In: *Information* 14.8, p. 426. DOI: https://doi.org/10.3390/info14080426

2. Michael Mayowa Farayola, Tal Irina, et al. (2024). "Towards Trustworthy AI: Potential and Peril of Integrating Multi-Phase Bias Mitigation Techniques in Recidivism Models". In: *Artificial Intelligence*, p. 104394

These publications reflect the main contributions of this chapter and provide further technical details, extended results, and peer-reviewed validation of the methods and findings.

Chapter 4

Methodology: Multi-Phase Fairness Pipelines and Optimization Techniques

4.1 Introduction

This chapter presents the methodological framework that underpins the empirical analyses of this thesis. Whereas Chapters 2 and 3 established the conceptual foundations and synthesized prior work on algorithmic fairness, this chapter translates those insights into a structured experimental design. This methodology was developed to address three central limitations in the literature: the fragmented application of fairness interventions, the persistence of fairness-accuracy trade-offs, and the limited generalizability of existing approaches across datasets and domains.

The framework is designed in response to **RQ2 (1.4): Can an integrated fairness-improving approach across different stages of AI development enhance fairness and predictive accuracy in recidivism prediction models?** Although this chapter does not provide the empirical results for RQ2 (1.4), it outlines the design decisions and reasoning that enable a systematic investigation. In particular, I integrate fairness-enhancing interventions across multiple stages of

the fairness pipeline, embed optimization techniques to capture fairness–accuracy trade-offs explicitly, and employ statistical significance testing to ensure that observed improvements are robust rather than due to random variation.

Together, these methodological choices establish the foundation for the empirical investigations reported in Chapters 5, 6, and 7.

4.2 Motivation and Problem Statement

Research on fairness of AI has typically applied interventions in isolation, for example through reweighing in pre-processing, adversarial debiasing in in-processing, or post-hoc adjustments in post-processing. Such fragmented approaches overlook the interdependencies within the machine learning pipeline, where bias may arise and propagate at multiple points. As a result, single-stage interventions often yield only partial or unstable improvements, highlighting the need for more holistic strategies (Caton and Haas 2023; Z. Chen et al. 2023).

A second unresolved challenge concerns the characterization of *fairness-accuracy trade-offs*. While many studies acknowledge that improving fairness can reduce predictive performance, few provide systematic mappings of these trade-offs across multiple fairness metrics. In high-stakes domains such as criminal justice, this lack of systematic evidence limits stakeholders' ability to make informed decisions. A more rigorous approach is required, one that makes explicit where compromises occur and identifies Pareto-efficient configurations in which fairness improvements can be achieved without disproportionate losses in accuracy(Caton and Haas 2023; Z. Chen et al. 2023).

A third limitation arises from the narrow range of benchmark datasets on which most integrated fairness-enhancing techniques are tested. Heavy reliance on a small set of datasets restricts understanding of how interventions behave under different demographic distributions, structural biases, or imbalance conditions. Without cross-domain validation, fairness-enhancing methods risk remaining confined to academic benchmarks rather than informing practice in diverse real-world settings.

Problem Statement: To address these limitations, we proposed and developed a structured methodological framework that integrates fairness-enhancing interventions across multiple pipeline stages and embeds optimization techniques to evaluate fairness-accuracy trade-offs across diverse datasets. This framework is designed not only to identify which configurations improve fairness, but also to assess their generalizability and stability across contexts (Caton and Haas 2023; Z. Chen et al. 2023).

By explicitly targeting fragmentation, trade-off characterization, and generalizability, this study establishes the rationale for constructing multi-phase fairness pipelines and embedding optimization-based evaluation as central components of the research methodology.

4.3 Overview of Methodological Approach

We developed the methodology around a multi-phase pipeline in which fairness-enhancing interventions are applied at the pre-processing, in-processing, and post-processing stages simultaneously. This structure was proposed to address biases that originate and compound at multiple points in the fairness pipeline, since interventions at only a single stage often produce limited or unstable gains. By enabling both isolated and integrated combinations of techniques, the framework supports a systematic investigation of how fairness outcomes are shaped by cumulative pipeline effects rather than by any single intervention.

To explicitly capture the inherent tensions between fairness and predictive accuracy, we embedded both many-objective (4.6.1) and bi-objective (4.6.4) optimization within the methodology. Many-objective optimization was employed to balance five objectives, accuracy (Acc), statistical parity difference (SPD), disparate impact (DI), equal opportunity difference (EOD), and predictive equality difference (PED), and to identify Pareto-efficient solutions. Bi-objective optimization was included to reflect practical decision contexts where one fairness metric is prioritized relative to accuracy. This dual use of optimization enables the methodology to uncover

not only the general landscape of fairness–accuracy trade-offs but also the targeted trade-offs that arise in real-world applications.

In addition, we incorporated statistical evaluation techniques, including standard deviation and confidence intervals, to determine whether observed improvements in fairness and accuracy were robust rather than artifacts of random variation. Embedding statistical significance analysis at the methodological level strengthens the reliability and interpretability of the results.

The framework is applied across multiple benchmark datasets in recidivism prediction, healthcare, and finance. These datasets were selected because they differ in demographic distributions, imbalance patterns, and embedded biases, thereby providing a robust test bed for assessing whether fairness interventions are generalizable or domain-specific. In doing so, the methodology positions fairness pipelines within diverse and challenging conditions rather than confining analysis to narrow benchmarks.

In summary, the purpose of this methodological chapter is not to provide direct answers to RQ2 (1.4), but to establish the structured experimental environment required to address it systematically in later chapters. By integrating fairness interventions across pipeline stages, embedding optimization-based evaluation, and applying statistical significance testing, the framework provides a principled foundation for the empirical analyses that follow. The next section introduces the datasets that form the empirical basis of the experiments.

4.4 Datasets Used

We evaluated the methodological framework using multiple benchmark datasets (criminal justice, finance) and real world dataset (health insurance). These datasets were deliberately selected not only because of their prevalence in fairness research but also because they embody distinct demographic distributions, imbalance characteristics, and embedded biases. By including datasets that differ substantially in structure and context, the study moves beyond narrow benchmarks and provides a

more rigorous test of whether fairness–performance trade-offs are stable or domain-specific.

The use of multiple datasets also reflects the overarching aim of this research: to design fairness-aware methodologies that are both technically robust and practically relevant. Criminal justice datasets such as COMPAS and RisCanvi provide canonical and real-world test cases where fairness concerns are well documented. The Adult Income dataset introduces a widely studied socioeconomic benchmark that enables comparability with prior work and evaluation outside the justice domain. Finally, the Irish health insurance dataset extends the analysis to a high-stakes domain where fairness is equally critical. Together, these datasets ensure that the empirical analyses are not confined to a single domain but instead reveal cross-contextual insights into fairness interventions.

The following subsections describe each dataset in detail, outlining its origin, structure, and specific relevance to the research objectives.

4.4.1 COMPAS Dataset

This study employs the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) dataset as a key benchmark for fairness research (Angwin et al. 2016a). The dataset contains 6,172 records from Broward County, Florida, with demographic and criminal history data across 11 features. The primary prediction target is two-year recidivism, defined as whether an individual reoffended within two years of release.

A filtered subset of 5,278 entries was used, restricted to individuals identified as either African American or Caucasian. This filtering decision reflects prior fairness research, which has consistently highlighted disparities in COMPAS outcomes between these two groups. Race was designated as the protected attribute, with Caucasians treated as the privileged group and African Americans as the unprivileged group. To extend the analysis, gender was also included as a sensitive attribute, enabling intersectional subgroup evaluation. This produced four subgroups:

Male Caucasian, Male African American, Female Caucasian, and Female African American. Male Caucasians were considered the privileged group, while the other three subgroups were treated as unprivileged, reflecting patterns reported in both the dataset and the wider literature.

The class distribution within the subset included 2,795 non-recidivists (label 0) and 2,483 recidivists (label 1), indicating a mild class imbalance. To prepare the dataset for fairness-aware models, categorical features were numerically encoded. Age was discretized into three categories: 0 ("less than 25"), 1 ("25–45"), and 2 ("greater than 45"), following established recidivism risk groupings. Recidivism risk was encoded as 0 ("Low"), 1 ("Medium"), and 2 ("High"). Gender was encoded as 0 ("Female") and 1 ("Male"), while race and charge severity were binarized with 0 for African American or misdemeanor and 1 for Caucasian or felony. These preprocessing steps ensured consistency across fairness interventions and comparability with prior studies.

4.4.2 RisCanvi Dataset

The RisCanvi dataset employed in this study is an anonymized and carefully curated resource derived from the fifth wave of a longitudinal research project on recidivism within the Catalonian prison system (Jurídics i Formació Especialitzada (CEJFE) 2020). Spanning more than three decades, the broader series has documented post-incarceration trajectories of released individuals. For the purposes of this research, I used the cohort of individuals released in 2015, who were tracked over a five-year follow-up period ending in December 2019. Data collection was intentionally concluded before March 2020 to avoid confounding effects from external disruptions such as the COVID-19 pandemic.

The dataset initially contained 3,814 individuals, categorized as final releases (67.4%), conditional releases (30.7%), suspended sentences (1.9%), and expulsion cases (3.1%). Because expelled individuals do not re-enter the Catalonian prison system, they were excluded from the recidivism analysis. Recidivism was defined as

penitentiary recidivism, including any new incarceration (pre-trial or post-conviction) during the observation window. This operationalization aligns with established definitions in criminal justice research while ensuring consistency with fairness-oriented evaluation.

A key motivation for including RisCanvi was its ability to capture region-specific dynamics within Catalonia's justice system, thereby complementing the more widely studied U.S.-based COMPAS dataset. Unlike COMPAS, RisCanvi incorporates both native Catalonian and non-native individuals, allowing fairness assessments to examine demographic disparities in a European context. For evaluation, native Catalonians were designated as the privileged group, while individuals from other regions were treated as the unprivileged group, enabling structured group-level comparisons.

Substantial preprocessing was undertaken to ensure data quality and usability. Missing values in yes/no/uncertain variables were imputed with "uncertain," while temporal imputation was applied where appropriate. Records with unresolved missing values were excluded, resulting in a final curated dataset of 2,885 individuals. Within this sample, 2,391 did not recidivate and 464 did, reflecting a moderate class imbalance. To mitigate this imbalance and support equitable subgroup evaluation, random sampling techniques were applied during experimentation.

I also contributed to the dataset's accessibility by conducting linguistic and structural adaptations. The original curated dataset was in Spanish; I translated it into English and standardized variable formats to ensure consistency across analytical tools. These adaptations facilitated integration into fairness-aware machine learning pipelines and improved reproducibility. To promote transparency and open science, the curated dataset and accompanying documentation are publicly available at: RisCanvi GitHub Repository.

4.4.3 Adult Income Dataset

The Adult Income dataset (Chakrabarty and S. Biswas 2018), also known as the Census Income dataset, originates from U.S. Census Bureau data and has become a

standard benchmark in algorithmic fairness research. It is particularly relevant for evaluating bias in socioeconomic status prediction, as the task involves classifying whether an individual's annual income exceeds $50,000 based on attributes such as age, education, marital status, occupation, and weekly working hours.

For this research, we employed a preprocessed subset of the dataset consisting of 48,842 records, filtered to include only individuals identified as either Black or White. This restriction reflects common practice in fairness studies, where disparities in predicted income between Black and White individuals are among the most well-documented sources of bias. To support intersectional fairness analysis, both race and gender were considered as sensitive attributes, resulting in four demographic subgroups: Male White, Female White, Male Black, and Female Black. Male White individuals were designated as the privileged group, while the remaining three subgroups were treated as unprivileged.

The classification problem is binary: individuals earning $50,000 or less were labeled as class 0, and those earning more than $50,000 were labeled as class 1. To prepare the dataset for fairness-aware models, categorical variables were transformed using one-hot or ordinal encoding, depending on feature type, while continuous features were normalized. These preprocessing steps were necessary to ensure comparability across methods and compatibility with fairness-enhancing techniques implemented in AIF360.

4.4.4 Irish Health Insurance Dataset

This study incorporates a private, sensitive real-world dataset sourced from a healthcare organization in Ireland. The dataset includes member-level information obtained through a combination of structured surveys, operational systems, and institutional metadata. It consists of 376,357 records with 177 features. The predictive task was to identify members likely to become "detractors", individuals providing negative satisfaction feedback, because detractor status serves as a practical proxy for perceived service quality in healthcare coverage. This outcome was chosen due

to its direct operational relevance and its potential consequences for both healthcare organizations and members.

The dataset contains both numerical and categorical variables, including protected attributes such as self-reported race/ethnicity, gender, and age. To enable intersectional fairness evaluation, these attributes were combined into a single composite feature, `RACE_GENDER_AGE`, resulting in ten defined subgroups (e.g., White Male 65+, Latino Female 65+, White Female under 65). Rare categories were aggregated into an "Other" class to maintain statistical validity. Each subgroup contained at least 9,217 records, ensuring adequate sample size for reliable fairness comparisons. This design choice reflected the need to balance detailed subgroup analysis with statistical robustness.

To mitigate risks of proxy discrimination, features highly correlated with protected or socioeconomic variables, such as regional census identifiers and prior-year quality scores, were excluded from training. Categorical features were encoded using one-hot encoding, while missing values were imputed using the mean for numerical attributes and the mode for categorical ones. These preprocessing steps were applied to ensure model compatibility, fairness validity, and consistency with best practices.

Strict ethical and privacy safeguards governed all data handling. Personal identifiers were fully anonymized, and variable names were withheld in accordance with the healthcare provider's policies and relevant data protection regulations. The dataset is not publicly accessible, and all analysis was conducted within a secure, ethically approved environment. Including this dataset allowed the methodology to be evaluated in a high-stakes, real-world healthcare context, where demographic diversity, incomplete data, and privacy constraints reflect the practical challenges of deploying fairness-aware AI systems.

Together, these four datasets, COMPAS, RisCanvi, Adult Income, and the Irish Health Insurance dataset, were selected to provide both breadth and depth in evaluating fairness interventions. They span distinct domains (criminal justice, healthcare, and socioeconomic prediction), geographic contexts (United States and Eu-

rope), and data scales (from thousands to hundreds of thousands of records). This diversity was intentional: it enables assessment of whether fairness-enhancing methods generalize across settings with different demographic distributions, structural biases, and imbalance characteristics. By grounding the experiments in datasets that reflect both benchmark standards and real-world complexity, the methodology ensures that findings are not confined to narrow academic cases but have broader applicability.

4.5 Experimental Setup

We evaluated fairness-enhancing interventions across the three standard stages of the machine learning pipeline, pre-processing, in-processing, and post-processing, by implementing four integrative configurations: PI (Pre-processing + In-processing), PP (Pre-processing + Post-processing), IP (In-processing + Post-processing), and PIP (Pre-processing + In-processing + Post-processing). These configurations were chosen to capture not only the effects of isolated interventions but also their cumulative and potentially synergistic impacts when combined. In this way, the experimental framework provides a structured basis for analyzing how fairness can be improved across the pipeline while maintaining predictive performance. Figure 4.1 illustrates the full fairness-enhanced pipeline structure.

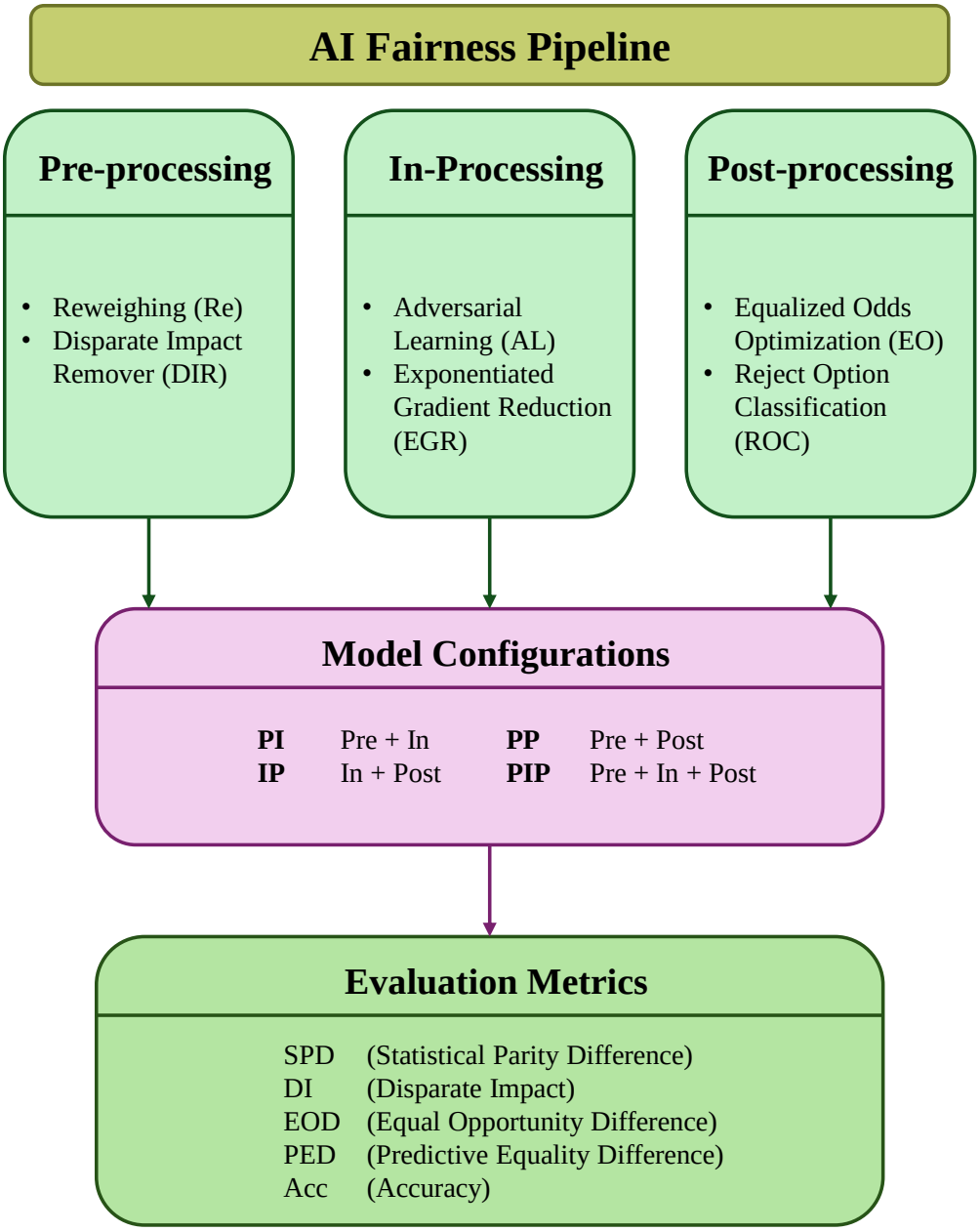

Figure 4.1: Fairness-enhancing techniques integrated across the AI Fairness Pipeline.

4.5.1 Datasets and Evaluation Configurations

Experiments were conducted on three datasets representing distinct domains: the COMPAS recidivism prediction dataset, the Adult Income dataset, and the Irish Health Insurance dataset. These datasets were selected because they differ in demographic attributes, imbalance characteristics, and embedded biases, thereby providing a diverse test bed for evaluating the generalizability of fairness interventions.

Each dataset was treated as a binary classification problem with protected attributes such as race, gender, or age depending on the domain context.

To preserve group comparability, we applied a 70:30 train–test split using stratified sampling, ensuring proportional representation of privileged and unprivileged groups across splits. To account for variability introduced by random initialization, each experimental setup was executed across 10 independent trials using different random seeds. Results were reported as averages across trials, with standard deviations and confidence intervals provided to quantify stability and robustness. Hyperparameters were held constant across pipeline configurations, except where fairness algorithms required specific parameters (e.g., adversarial loss weight, repair level). This standardized protocol ensured that observed differences could be attributed to fairness interventions rather than to sampling or model-training variance. The statistical procedures used to summarize and assess these results are detailed in Section 4.6.5.

All models were implemented using the AIF360 fairness toolkit, which provides standardized implementations of fairness methods and evaluation metrics. Fairness outcomes were computed using the `BinaryLabelDataset` structure to ensure consistency across datasets and configurations. This design choice guaranteed that results reflected methodological differences rather than inconsistencies in implementation.

4.5.2 Fairness Techniques Implementations

Six fairness-enhancing methods were implemented across the pipeline stages: Preprocessing (Reweighing (Re), Disparate Impact Remover (DIR)), In-processing (Adversarial Learning (AL), Exponentiated Gradient Reduction (EGR)), and Postprocessing (Equalized Odds Optimization (EO), Reject Option-Based Classification (ROC)). These methods were selected because they represent widely studied and complementary approaches to fairness: pre-processing techniques directly address bias in data distributions, in-processing techniques embed fairness constraints during model training, and post-processing techniques adjust decision thresholds when

retraining is not feasible. By covering all three stages, the study ensures that the methodology captures both isolated and cumulative fairness effects.

Each method was integrated into one or more pipeline configurations (PI, PP, IP, or PIP), enabling systematic evaluation of how different combinations influence fairness and accuracy. This integrative design allows the analysis to go beyond individual methods and to assess how fairness interventions interact when combined, revealing both trade-offs and potential synergies.

All implementations were carried out using the `AIF360` fairness library, supported by `scikit-learn` and `TensorFlow`. The use of `AIF360` ensured standardized implementations of fairness techniques and metrics, while complementary libraries provided flexibility for model development. To support transparency and reproducibility, the complete source code is openly available at: Integrated Fairness Techniques.

Table 4.1 summarizes these relationships, showing how each fairness-enhancing method aligns with the fairness metrics it most directly influences.

Table 4.1: Fairness-Improving Methods Vs. Corresponding Metrics Influence

		SPD	DI	EOD	PED
Pre-processing	Re	✓	✓		
	DIR	✓	✓		
In-processing	EGR	✓		✓	✓
	AL	✓	✓	✓	✓
Post-processing	ROC	✓	✓	✓	✓
	EO			✓	✓

As shown in Table 4.1, the methods span complementary objectives across pre-, in-, and post-processing stages. This mapping guided the experimental design by clarifying which fairness dimensions were most likely to be affected by each intervention.

The mapping of fairness-enhancing methods to their primary metric influences in this study is not arbitrary but grounded in findings from state-of-the-art literature Bellamy et al. 2019. Prior comparative analyses and benchmark studies have documented the conditions under which particular techniques tend to improve spe-

cific fairness measures (Bellamy et al. 2019; M. Zhang and Sun 2022; Mehrabi et al. 2021). Building on these insights, we structured the methodological framework so that each intervention is linked to the fairness metrics it most directly influences, ensuring that the design is both evidence-based and aligned with established findings.

For example, pre-processing methods such as Reweighing and Disparate Impact Remover have consistently been shown to reduce Statistical Parity Difference (SPD) and improve Disparate Impact (DI). In-processing approaches such as Exponentiated Gradient Reduction and Adversarial Learning operate during training to embed fairness constraints, influencing both Equal Opportunity Difference (EOD) and Predictive Equality Difference (PED). Post-processing strategies, including Equalized Odds and Reject Option-Based Classification, adjust decision thresholds and are therefore particularly effective at improving group-level parity in true and false positive rates.

Table 4.1 synthesizes these findings by categorizing fairness methods into pre-processing, in-processing, and post-processing approaches, and linking them to the fairness metrics most directly affected. While each method is primarily associated with one or two fairness objectives, secondary effects often extend to other metrics due to the inherent interdependencies among fairness definitions. In this way, the table serves both as a design reference for the experimental pipeline and as a reminder of the potential ripple effects that fairness interventions can introduce.

Pre-processing Techniques

Reweighing (Re). We applied Reweighing prior to model training to adjust for representational bias across groups. The algorithm increases the weights of underrepresented favorable outcomes from the unprivileged group and decreases the weights of overrepresented outcomes from the privileged group. These weights are then carried through to the classifier during training, effectively simulating a more balanced distribution without altering the dataset structure. Reweighing was se-

lected because it addresses disparities in outcome representation while preserving the original feature space, thereby minimizing disruption to downstream processing and allowing direct comparison with unmodified datasets.

Disparate Impact Remover (DIR). We implemented the Disparate Impact Remover with a repair level of 1.0, corresponding to full transformation of features correlated with the protected attribute. This configuration was chosen to maximize the removal of biased correlations while retaining the relative ranking of individuals, ensuring that predictive utility was not discarded entirely. Using a full repair level also enabled this study to isolate the effects of complete de-biasing, providing a clear comparison with scenarios involving partial transformations. The transformed dataset was then incorporated into downstream stages depending on the pipeline configuration: in PI and PIP models, the debiased features were used as inputs to in-processing methods such as Exponentiated Gradient Reduction, while in PP models, the transformed dataset was passed to logistic regression classifiers before proceeding to post-processing interventions. This design ensured that the role of DIR could be systematically assessed both in isolation and in combination with other fairness-enhancing techniques.

Note on Oversampling

Although oversampling methods such as Random Over-Sampling and SMOTE are commonly used as pre-processing techniques to address class imbalance, they were not included in the integrated fairness-enhancing framework (PI, PP, IP, PIP) described in this chapter. This methodological decision was deliberate. Oversampling primarily seeks to rebalance class distributions rather than directly enforce fairness constraints, and its effects on fairness can be complex and dataset-dependent. In particular, oversampling may unintentionally introduce synthetic bias or lead to overfitting when sensitive attributes are correlated with target labels.

To maintain the conceptual clarity of the integrated pipeline, oversampling was therefore excluded from the main framework and investigated separately in Chap-

ter 6. This separation allowed for a more focused examination of the trade-offs introduced by oversampling, while ensuring that the fairness pipeline (pre-, in-, and post-processing interventions) remained centered on explicit fairness-enhancing methods. Together, the two analyses provide complementary insights into how fairness can be improved through both integrated interventions and targeted imbalance mitigation.

In-processing Techniques

Adversarial Learning (AL). We implemented Adversarial Debiasing using a neural network as the primary predictor, with an adversarial component simultaneously trained to infer the protected attribute from the predictor's outputs. The model was trained for 50 epochs with a batch size of 128, settings selected to ensure stable convergence across datasets while supporting generalization. The adversarial loss weight was fixed at 0.2 to balance predictive performance with the strength of fairness enforcement. Enabling the `debias` parameter ensured that the predictor actively suppressed protected attribute signals. This configuration was chosen because adversarial learning directly targets the leakage of protected information into predictions, thereby addressing bias at its source during training.

Exponentiated Gradient Reduction (EGR). We configured Exponentiated Gradient Reduction with logistic regression as the base classifier. The fairness constraint was set to `DemographicParity`, requiring that positive outcomes be equally distributed across groups. This constraint was chosen because it enforces group-level parity while being computationally tractable for integration with multiple pipeline configurations. Training was limited to a maximum of 20 iterations to balance the need for convergence with computational efficiency. The algorithm produces a randomized ensemble of classifiers, weighted to minimize classification error while simultaneously enforcing fairness. Retaining the protected attribute during training allowed fairness effects to be explicitly quantified during post-training evaluation. This setup ensured that EGR's ability to reconcile accuracy with fairness could be

systematically compared with both adversarial learning and other pipeline interventions.

Post-processing Techniques

Equalized Odds Optimization (EO). We applied Equalized Odds Optimization to model predictions after training. Calibration probabilities were estimated from a validation set and used to guide label adjustments so that true positive and false positive rates were aligned across groups. This technique was selected because it specifically addresses fairness in error distributions, measured by *Equal Opportunity Difference (EOD)* and *Predictive Equality Difference (PED)*. The rationale for including EO was its ability to impose fairness constraints exclusively at the output stage, without requiring model retraining. This made it particularly useful for evaluating fairness gains that arise solely from post-processing adjustments.

Reject Option-Based Classification (ROC). We implemented Reject Option-Based Classification on instances near the decision boundary, where the model exhibited low confidence. In these cases, favorable outcomes were preferentially assigned to unprivileged groups, while unfavorable outcomes were assigned to privileged groups. Decision thresholds were explored in the range 0.3–0.8, with fairness bounds set to ± 0.05 for Equal Opportunity Difference. These parameters were determined through grid search, allowing a balance between fairness improvements and predictive performance. ROC was chosen because it enables targeted fairness adjustments in borderline cases, providing a practical mechanism for post-deployment correction when retraining is either infeasible or undesirable.

4.5.3 Model Integration and Metric Computation

We constructed each pipeline configuration (PI, PP, IP, PIP) by systematically combining the respective techniques from the pre-, in-, and post-processing stages. Models were then evaluated using five metrics: Accuracy (reported both in aggregate and by subgroup), Statistical Parity Difference (SPD), Disparate Impact (DI),

Equal Opportunity Difference (EOD), and Predictive Equality Difference (PED). These metrics were chosen because they represent complementary perspectives on fairness, ranging from group-level parity (SPD, DI) to fairness in error distributions (EOD, PED), and together provide a more comprehensive assessment than any single measure.

The choice of these four fairness metrics, Statistical Parity Difference (SPD), Disparate Impact (DI), Equal Opportunity Difference (EOD), and Predictive Equality Difference (PED), was deliberate. They represent complementary perspectives on algorithmic fairness: SPD and DI capture group-level parity in positive outcomes, while EOD and PED focus on parity in error distributions (true positive and false positive rates, respectively). This combination reflects both allocation fairness and error fairness, providing a balanced view of potential harms. By focusing on SPD, DI, EOD, and PED, the methodology aligns with widely adopted standards in fairness research, especially in criminal justice system, finance and health sector.

Fairness metrics were computed both in aggregate (privileged versus all unprivileged groups combined) and at the subgroup level (e.g., Black Female, White Male). This design was motivated by the recognition that aggregate analyses may obscure intersectional disparities, whereas subgroup reporting makes it possible to capture fine-grained patterns of bias. Subgroup accuracy was also reported to identify whether fairness interventions produced unintended performance degradation in specific demographic subgroups.

All fairness-enhanced models were trained and evaluated under uniform conditions, using consistent data splits, hyperparameters, and evaluation protocols. This controlled setup was necessary to isolate the effects of fairness interventions and to ensure that observed differences were attributable to the interventions themselves rather than to confounding design variations. After training, predictions were analyzed across all metrics to identify which pipeline configurations most effectively balanced fairness and predictive accuracy.

Classifier selection was also made with methodological intent. Logistic regression

was employed in several configurations because it is interpretable, computationally efficient, and widely used as a baseline in fairness research. Its linear structure allows fairness interventions such as Exponentiated Gradient Reduction to be directly integrated with clear interpretability of outcomes. Neural networks were used for Adversarial Debiasing, as this approach requires a flexible model architecture to jointly optimize predictions and fairness constraints via adversarial training. These choices ensured that the fairness techniques could be evaluated in line with their canonical implementations in AIF360 while keeping results transparent and comparable with prior literature. More complex model families were not included in the core framework to maintain clarity and focus on fairness interventions rather than on model-specific effects.

All experiments and metrics were implemented using reproducible scripts in the `AIF360` library, with supporting models from `scikit-learn` and `TensorFlow`. The use of AIF360 provided standardized implementations of fairness interventions and metrics, enabling comparability across methods. To support transparency and open science, detailed logs and scripts are available in the public repository: https://github.com/mmfara/IF-EA.

4.6 Multi-Objective Optimization Framework and Statistical Significance

Selecting the most appropriate model from among the PI, IP, PP, and PIP configurations requires balancing fairness with predictive accuracy. Because there is no universally accepted fairness metric and fairness definitions often conflict with accuracy, we adopted a multi-objective optimization framework to guide model selection. This approach allows systematic exploration of fairness–accuracy trade-offs, rather than privileging any single metric. In addition, we evaluated the statistical significance of observed improvements across both accuracy and fairness metrics to ensure that results were robust and not due to random variation.

4.6.1 Multi-Objective Optimization and the Pareto Front

Multi-objective optimization enables the simultaneous evaluation of competing objectives without reducing them to a single composite score. In this study, the method was used to identify non-dominated solutions among the candidate models (PI, IP, PP, and PIP), where a solution is considered non-dominated if it performs at least as well as all others across every objective and strictly better on at least one. This design choice was made to preserve the inherent complexity of fairness evaluation, where improvements in one measure may worsen another.

The analysis is grounded in the concept of the Pareto front (Sharma and Kumar 2022), which represents the set of optimal trade-offs: improving one objective along the front necessarily entails compromising another. By focusing on Pareto-optimal models, this study provides a flexible framework for context-specific model selection, rather than relying on an arbitrary fairness or accuracy threshold.

4.6.2 Formulating the Objectives

In constructing the optimization framework, we treated predictive accuracy as an objective to be maximized, while fairness metrics required normalization to be expressed consistently in the same optimization space. Metrics such as Statistical Parity Difference (SPD), Equal Opportunity Difference (EOD), and Predictive Equality Difference (PED) are ideal when their values approach zero, whereas Disparate Impact (DI) is optimal when approaching one. To ensure comparability, the absolute values of SPD, EOD, and PED were minimized to penalize deviations from zero, while DI was transformed inversely when it fell below one to encourage convergence toward the ideal value of one. This normalization step allowed all fairness objectives to be framed under a minimization formulation, ensuring consistency and enabling balanced comparison with accuracy in the multi-objective setting.

4.6.3 Non-Dominated Solutions and Pareto Analysis

Let $O = [O_1(x), \ldots, O_k(x)]$ represent the vector of k objective values (e.g., Accuracy, SPD, DI, EOD, PED) for a model x. A model x_1 is said to dominate another model x_2, denoted $x_1 \succ x_2$, if:

$$\forall i \in \{1, \ldots, k\},\ O_i(x_1) \geq O_i(x_2) \quad \text{and} \quad \exists j \in \{1, \ldots, k\} \text{ such that } O_j(x_1) > O_j(x_2)$$

A solution is non-dominated if no other model dominates it. The set of all such non-dominated solutions forms the Pareto front. Each model on the Pareto front represents an optimal balance between accuracy and fairness, where no further improvement can be made in one objective without compromising another. This analytical approach was chosen to make explicit the trade-offs that practitioners and policymakers must consider when selecting models for deployment.

4.6.4 Bi-Objective Optimization for Prioritized Fairness

Although many-objective optimization provides a comprehensive landscape of trade-offs, it can complicate interpretation, particularly when multiple fairness metrics are considered simultaneously. In real-world settings, however, stakeholders often prioritize one fairness criterion over others depending on regulatory, ethical, or policy contexts. To reflect this reality, I also employed a bi-objective optimization framework that pairs accuracy with a single fairness metric (e.g., SPD, EOD, or PED).

This complementary analysis simplifies interpretation and visualization while enabling targeted model selection based on specific fairness goals (Schwind et al. 2020; Sharma and Kumar 2022). By evaluating each fairness–accuracy pair individually, the framework offers a practical means of balancing predictive performance with fairness in alignment with domain-specific priorities.

4.6.5 Statistical Evaluation

We employed statistical analysis to ensure that the experimental results were both reliable and interpretable. Since fairness interventions often yield variable outcomes across different runs due to random initialization and sampling, statistical evaluation was necessary to assess whether observed differences reflected systematic effects rather than noise. For each metric, results were summarized using the mean, standard deviation, and 95% confidence interval.

The mean was used to provide a central estimate of performance or fairness across multiple experimental trials, offering a representative value for comparison across models. The standard deviation (std) quantified the dispersion of results around the mean, thereby indicating the consistency and stability of each fairness intervention. The 95% confidence interval (CI) defined the statistical range within which the true value of a metric is expected to lie with 95% confidence, providing an indication of the precision of the estimated outcomes. These measures were selected because together they balance interpretability with statistical rigor, enabling a robust evaluation of both fairness and accuracy across experimental configurations.

By incorporating these statistical measures, the study not only reports point estimates but also accounts for variability and uncertainty, strengthening the transparency and robustness of the methodological framework.

4.7 Ethical Considerations

The design and implementation of this research were grounded in ethical principles intended to promote fairness, transparency, and social responsibility in artificial intelligence systems. Because the study focuses on high-stakes domains such as criminal justice and healthcare, ethical rigor was a priority at every stage of the project.

All datasets used in this research, COMPAS, RisCanvi, Adult Income, and the Irish Health Insurance dataset, were either publicly available or accessed under strict

ethical guidelines. Personally identifiable information (PII) was absent or removed prior to analysis, and all data handling complied with established research ethics protocols. These safeguards ensured that the work did not involve interventions on individuals or real-time decision-making systems, thereby eliminating risks of direct harm.

The decision to implement fairness interventions was motivated by the ethical imperative to reduce harm for marginalized populations, particularly those affected by compound or intersectional bias. Throughout the study, I was attentive to the risk that algorithmic models could reinforce existing systemic inequalities if not critically evaluated and constrained.

The choice of fairness metrics, such as statistical parity difference and equal opportunity difference, and trade-off analysis methods, including Pareto front optimization, was informed not only by their statistical rigor but also by their ethical relevance in evaluating disparate impacts across demographic groups. These methodological choices ensured that fairness was assessed in ways that are meaningful for real-world stakeholders.

Finally, we promoted ethical transparency through the use of open-source tools, reproducible pipelines, and detailed documentation. By making the methods and code publicly available, the study supports accountability, enables independent scrutiny, and contributes to a more transparent and socially responsible approach to fairness in machine learning.

4.8 Limitations of Methodology

While the methodological framework developed in this thesis is robust and systematically designed, several limitations must be acknowledged.

First, the analysis is restricted to tabular datasets and binary classification tasks. This scope was chosen to enable controlled evaluation across multiple domains, but it excludes other important data modalities such as text, images, and multi-label outputs, which may display different fairness dynamics. Future work should extend

the framework to these data types.

Second, although the pipeline integrates pre-, in-, and post-processing mitigation techniques, it does not model real-world decision feedback loops or long-term societal impacts. These dynamics are particularly important in domains such as recidivism and healthcare, where algorithmic predictions may influence future data generation. The focus here was on isolating and systematically evaluating fairness interventions, with broader feedback effects left as a direction for subsequent research.

Third, the evaluation of fairness is constrained by the current state of fairness definitions, many of which are mathematically incompatible and may not fully reflect lived experiences of unfairness. While subgroup analysis was included to capture intersectional disparities, the statistical reliability of results was limited by small sample sizes in certain categories. This reflects a broader challenge in fairness research rather than a methodological oversight.

Fourth, while oversampling was discussed in Chapter 6 as a major investigation, it was not incorporated into the integrated fairness-enhancing framework presented in Chapter 4. This separation was intentional: oversampling primarily addresses class imbalance rather than fairness constraints, and it may introduce synthetic bias or overfitting when applied with sensitive attributes.

Finally, although the inclusion of datasets from criminal justice, healthcare, and socioeconomic domains improves generalizability, the findings cannot be assumed to transfer universally. Cultural, legal, and systemic differences across jurisdictions may alter both how bias manifests and how effective mitigation strategies prove in practice.

Despite these limitations, the methodology provides a structured and transparent empirical foundation for evaluating fairness-aware AI systems. By explicitly acknowledging these constraints, this study highlights important areas for future exploration while reinforcing the value of the current contributions.

4.9 Conclusion

This chapter has outlined the methodological framework that underpins the empirical analyses of this thesis. The framework was designed in direct response to limitations identified in the literature: the fragmented application of fairness interventions, the persistence of fairness–accuracy trade-offs, and the lack of generalizability across domains. To address these challenges, we developed a multi-phase fairness pipeline that integrates pre-, in-, and post-processing techniques in both isolated and combined configurations (PI, PP, IP, PIP). This structure ensures that fairness can be studied not only as the effect of single interventions but also as the cumulative outcome of multiple stages in the machine learning lifecycle.

The experimental setup was applied to four diverse datasets spanning criminal justice (COMPAS, RisCanvi), socioeconomic prediction (Adult Income), and healthcare (Irish Health Insurance). These datasets differ in demographic distributions, imbalance patterns, and embedded biases, providing a rigorous test bed for assessing both the stability and the generalizability of fairness interventions. A standardized evaluation protocol was followed, including stratified train–test splits, multiple trials with varying random seeds, and consistent hyperparameters, with results summarized using statistical significance testing to ensure robustness and interpretability.

The methodology further embedded a multi-objective optimization framework, leveraging Pareto analysis to capture the inherent tensions between fairness and predictive accuracy. Both many-objective and bi-objective formulations were employed, enabling analysis of general trade-offs as well as targeted scenarios in which a single fairness metric is prioritized. By aligning fairness evaluation with optimization techniques, the framework enables transparent and principled model selection.

Ethical considerations and methodological limitations were also explicitly addressed. The use of fairness metrics was justified on the basis of their complementarity and their alignment with widely recognized standards. All experiments were conducted with attention to privacy, and transparency, while acknowledging con-

straints such as dataset scope, the binary classification focus, and the absence of real-world feedback loops. These limitations frame directions for future research while underscoring the rigor of the present design.

In summary, this methodological chapter establishes a principled and reproducible foundation for the analyses that follow. By combining fairness-aware pipelines, optimization-based evaluation, statistical rigor, and ethical responsibility, it positions the thesis to systematically address RQ2 in the empirical chapters that follow.

Chapter 5

Fairness Pipeline Integration: Technical and Empirical Investigations

5.1 Introduction

This chapter investigates the empirical performance of multi-phase fairness pipelines in the context of recidivism prediction. While Chapter 4 outlined the methodological foundations of fairness integration and optimization, the present chapter applies these methods to real-world datasets to assess their effectiveness. The focus here is on systematically evaluating combinations of fairness mitigation strategies across pre-processing, in-processing, and post-processing stages, with the aim of understanding how integrated approaches influence both fairness and predictive accuracy.

The motivation for this investigation stems from the limitations of single-phase interventions, which often improve one fairness dimension at the expense of others or lead to notable accuracy losses. By contrast, multi-phase integration has the potential to leverage the complementary strengths of different mitigation techniques, thereby reducing systemic disparities more holistically. Through comparative analysis on two benchmark datasets, COMPAS and RisCanvi, we examine not only the

fairness-accuracy trade-offs within each pipeline but also the broader implications of integrating bias mitigation across the AI lifecycle.

The objective of this chapter is therefore threefold: (i) to evaluate the empirical outcomes of fairness pipeline integrations across multiple configurations, (ii) to analyze the trade-offs between fairness metrics and accuracy under many-objective and bi-objective optimization, and (iii) to identify key patterns, insights, and dataset-specific behaviors that inform the design of trustworthy and equitable recidivism prediction models.

5.1.1 Research Questions and Hypothesis

This research contributes to the evolving field of algorithmic fairness by investigating the balance between equity and predictive performance, as highlighted in prior work S. Liu and Vicente 2022; Plecko and Bareinboim 2024; Mehrabi et al. 2021. Specifically, the study explores the effectiveness and scalability of a multi-phase fairness-enhancing pipeline that incorporates interventions at the pre-processing, in-processing, and post-processing levels. Emphasis is placed on validating this approach within the context of recidivism prediction, particularly using the RisCanvi dataset, which presents substantial real-world complexity and imbalance. Such validation is critical for assessing the resilience and practicality of fairness methods in domains where decisions carry significant societal consequences, such as criminal justice.

The core hypothesis asserts that implementing fairness interventions across multiple stages of the AI lifecycle can improve fairness outcomes while preserving classification accuracy, even when applied to heterogeneous and imbalanced datasets. To evaluate this hypothesis, the study is guided by the following sub-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, 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?

5.2 Results and Comparative Analysis

This section presents the empirical results obtained from applying the multi-phase fairness pipelines to the COMPAS and RisCanvi datasets, following the methodological framework described in Chapter 4. We begin by reporting outcomes for each integration category, PI (Pre-processing + In-processing), IP (In-processing + Post-processing), PP (Pre-processing + Post-processing), and PIP (Pre-processing + In-processing + Post-processing), to evaluate their comparative performance across fairness and accuracy metrics. For consistency, results are interpreted with respect to the ideal values of each fairness metric (SPD, EOD, and PED $\rightarrow 0$, DI $\rightarrow 1$) and accuracy maximization. In addition, standardized values (marked with an asterisk in Table 5.1) are discussed where minimization transformations were applied. The comparative analysis highlights not only the absolute performance of each model but also the trade-offs introduced when integrating fairness-enhancing techniques across different phases of the pipeline.

Table 5.1: 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

5.2.1 PI Models

The PI models, coloured in blue in table 5.1, which combine pre-processing and in-processing techniques, include `Re+EGR`, `DIR+EGR`, `Re+AL`, and `DIR+AL`. Their performance is compared across both datasets using the fairness metrics (SPD, DI, EOD, PED) and accuracy.

In terms of Statistical Parity Difference (SPD), COMPAS achieves its lowest disparity with `DIR+AL` (0.009), which is notably close to the ideal of 0. Other COMPAS PI models, such as `Re+EGR` (0.033) and `Re+AL` (0.061), also perform moderately well, while `DIR+EGR` (0.055) remains competitive. For RisCanvi, the lowest SPD is recorded by `Re+EGR` (0.058), followed by `Re+AL` (0.064). By contrast, `DIR+EGR` (0.135) represents the largest deviation. This indicates that COMPAS PI models

generally perform better in reducing parity disparities, though RisCanvi also yields some near-ideal values under specific interventions.

With respect to Disparate Impact (DI), where the ideal value is 1 and starred values (*) denote standardized inversions, COMPAS again performs strongly. `DIR+AL` (1.020*) is closest to parity, while `Re+EGR` (1.086*) and `DIR+EGR` (1.142*) also remain near the target. In RisCanvi, however, the results are more mixed: `Re+EGR` (1.189) and `Re+AL` (1.238*) perform moderately, whereas `DIR+EGR` (1.536*) represents the largest deviation across all PI models. This suggests that DI outcomes in RisCanvi are more variable and sensitive to the choice of integration.

For Equal Opportunity Difference (EOD), COMPAS PI models show relatively small magnitudes, with `Re+EGR` (0.026) and `DIR+AL` (0.028) performing best. `Re+AL` (0.052) and `DIR+EGR` (0.038) also remain fairly close to the ideal. RisCanvi shows greater variability, with `Re+AL` (0.059) and `Re+EGR` (0.068) performing reasonably well, while `DIR+EGR` records a much larger disparity (0.169). This indicates that true positive rate equality is harder to stabilize in RisCanvi under PI interventions.

Turning to Predictive Equality Difference (PED), COMPAS once again records low magnitudes, with `Re+AL` (0.004) and `DIR+EGR` (0.004) achieving near-ideal results, while `Re+EGR` (0.032) and `DIR+AL` (0.067) remain small. RisCanvi displays more variation: `Re+AL` (0.055) and `DIR+AL` (0.0640) are moderate, but `DIR+EGR` (0.142) deviates considerably. This suggests that false positive rate disparities are more challenging to mitigate effectively in RisCanvi.

In terms of accuracy, COMPAS PI models are consistent in the 0.66–0.67 range, while RisCanvi achieves slightly higher values between 0.71–0.73. Notably, `DIR+EGR` attains the highest RisCanvi accuracy (0.73), though it also corresponds to the largest fairness disparities, highlighting a clear trade-off.

Comparative Insights. The PI models demonstrate the potential and challenges of combining pre- and in-processing techniques. For COMPAS, interventions such as `DIR+AL` and `Re+EGR` yield consistently strong fairness outcomes across multiple metrics, albeit at modest accuracy levels. RisCanvi, in contrast, achieves higher

predictive accuracy, but exhibits wider variability in fairness outcomes, with models such as DIR+EGR amplifying disparities even as accuracy improves. These findings underscore the fairness–accuracy trade-offs inherent in PI integrations: while certain combinations achieve near-ideal fairness without major accuracy loss, others demonstrate that integration alone does not guarantee improvement and must be carefully assessed in a dataset-specific context.

5.2.2 PP Models

The PP models, coloured in pink in table 5.1, combine pre-processing with post-processing interventions and include Re+EO, Re+ROC, DIR+EO, and DIR+ROC. Their comparative performance across COMPAS and RisCanvi highlights how such integrations influence both fairness and predictive accuracy.

In terms of Statistical Parity Difference (SPD), RisCanvi performs strongly under Re+EO (0.010) and DIR+EO (0.010), which are very close to the ideal of 0. DIR+ROC (0.023) remains competitive, while Re+ROC (0.086) reflects a larger deviation. For COMPAS, SPD is lowest under DIR+EO (0.032) and Re+EO (0.034), while both DIR+ROC (0.288) and Re+ROC (0.081) produce considerably higher disparities. These results suggest that EO-based combinations are more effective than ROC-based ones in reducing group-level disparities.

For Disparate Impact (DI), COMPAS achieves near-ideal performance under DIR+EO (1.097) and Re+EO (1.095), whereas DIR+ROC (2.160) produces a substantial deviation. RisCanvi similarly records strong results under Re+EO (1.035*) and DIR+EO (1.037*), while Re+ROC (1.340) and DIR+ROC (1.086) diverge further from the ideal. Overall, EO-based integrations consistently improve DI, while ROC-based models tend to amplify disparities.

Turning to Equal Opportunity Difference (EOD), RisCanvi achieves its best results with DIR+EO (0.007) and Re+EO (0.020), which are very close to zero, while Re+ROC (0.092) represents the largest deviation. For COMPAS, Re+EO (0.001) and DIR+EO (0.007) again stand out as near-ideal, while DIR+ROC (0.300) produces one

of the highest disparities among all PP models. These findings reinforce the observation that EO-driven post-processing effectively improves true positive rate equity, whereas ROC-driven post-processing can worsen it.

With respect to Predictive Equality Difference (PED), RisCanvi achieves its lowest values under `DIR+EO` (0.0001) and `Re+EO` (0.002), both essentially ideal. In contrast, `Re+ROC` (0.075) and `DIR+ROC` (0.011) produce larger disparities. COMPAS mirrors this pattern: `Re+EO` (0.002) and `DIR+EO` (0.005) achieve near-zero PED, while `Re+ROC` (0.002) remains small, but `DIR+ROC` (0.207) is considerably higher. Once again, EO-based approaches provide superior stability in mitigating false positive rate disparities.

Regarding accuracy, RisCanvi PP models perform consistently well, ranging from 0.70–0.74, with `DIR+EO` and `DIR+ROC` achieving the highest accuracy (0.74). COMPAS PP models, by contrast, operate within a lower range of 0.64–0.68. This suggests that fairness improvements in COMPAS through post-processing often come with greater accuracy costs compared to RisCanvi.

Comparative Insights. The PP models reveal a clear distinction between EO- and ROC-based integrations. EO-based approaches (`Re+EO`, `DIR+EO`) consistently yield near-ideal fairness results across SPD, DI, EOD, and PED in both datasets, but especially in RisCanvi where these models achieve simultaneously high accuracy and strong fairness. By contrast, ROC-based integrations (`Re+ROC`, `DIR+ROC`) generally amplify disparities even when accuracy is stable or improved. These findings highlight that the effectiveness of post-processing depends strongly on the method chosen: EO-based interventions complement pre-processing well, while ROC-based combinations risk undermining fairness gains. The fairness–accuracy trade-off is thus shaped not only by the dataset but also by the specific type of post-processing integrated.

5.2.3 IP Models

The IP models, coloured in green in table 5.1, integrate in-processing with post-processing methods, specifically `EGR+EO`, `EGR+ROC`, `AL+EO`, and `AL+ROC`. Their evaluation across COMPAS and RisCanvi demonstrates how such combinations influence fairness and predictive performance.

In terms of Statistical Parity Difference (SPD), RisCanvi achieves very low disparities with `EGR+EO` (0.009) and `AL+EO` (0.011), both near the ideal value of zero. By contrast, `EGR+ROC` (0.021) and `AL+ROC` (0.024) produce slightly higher values, though still modest. COMPAS follows a similar trend: `EGR+ROC` (0.023) yields the lowest SPD, followed closely by `AL+EO` (0.036) and `EGR+EO` (0.036). `AL+ROC` (0.079) shows a larger deviation. These results suggest that EO-based post-processing produces stronger parity outcomes compared to ROC-based combinations, though both remain generally effective.

For Disparate Impact (DI), COMPAS achieves the closest-to-ideal results under `EGR+ROC` (1.059) and `AL+ROC` (1.159), while `EGR+EO` (1.095) and `AL+EO` (1.099) remain close to parity. RisCanvi shows a similar pattern: `EGR+EO` (1.024*) is near-ideal, while `AL+EO` (1.036*), `EGR+ROC` (1.066), and `AL+ROC` (1.092*) remain slightly further but still competitive. Overall, DI results indicate that both EO- and ROC-based IP integrations deliver relatively balanced outcomes, with RisCanvi showing slightly stronger proximity to the ideal.

Turning to Equal Opportunity Difference (EOD), COMPAS exhibits near-zero disparities with `EGR+EO` (0.003) and `AL+EO` (0.014), while `EGR+ROC` (0.015) also remains small. `AL+ROC` (0.037) is somewhat higher but still moderate. RisCanvi mirrors this: `EGR+EO` (0.008) and `AL+EO` (0.015) yield strong fairness, while `EGR+ROC` (0.011) and `AL+ROC` (0.028) remain close to the ideal. Thus, IP models generally excel in reducing true positive rate disparities across both datasets.

With respect to Predictive Equality Difference (PED), RisCanvi achieves near-ideal results under `EGR+EO` (0.001) and `AL+EO` (0.0002), while `EGR+ROC` (0.034) and `AL+ROC` (0.013) remain small but less optimal. COMPAS shows similar behavior:

EGR+EO (0.003) and AL+EO (0.016) perform strongly, while EGR+ROC (0.042) and AL+ROC (0.033) are slightly higher. These outcomes suggest that EO-based combinations provide more consistent improvements in controlling false positive rate disparities.

In terms of accuracy, RisCanvi models span 0.67–0.74, with AL+ROC (0.74) achieving the highest. COMPAS IP models fall within 0.65–0.68, with AL+ROC again attaining the highest (0.68). This indicates that ROC-based combinations can enhance accuracy more than EO-based ones, but often at the expense of fairness.

Comparative Insights. The IP models demonstrate that integrating in-processing with post-processing can deliver near-ideal fairness across multiple metrics, particularly with EO-based interventions. Both datasets show strong mitigation of disparities, though COMPAS records slightly lower fairness magnitudes overall. ROC-based combinations provide accuracy advantages (e.g., AL+ROC), but this often comes with higher disparities, reinforcing the fairness–accuracy trade-off. In contrast, EO-based integrations (EGR+EO, AL+EO) provide robust fairness improvements with minimal accuracy loss, making them more reliable for fairness-sensitive applications.

5.2.4 PIP Models

The PIP models, coloured in yellow in table 5.1, represent the most comprehensive integration, combining pre-processing, in-processing, and post-processing techniques. This group includes Re+EGR+EO, Re+EGR+ROC, Re+AL+ROC, Re+AL+EO, DIR+EGR+ROC, DIR+EGR+EO, DIR+AL+ROC, and DIR+AL+EO. Their evaluation highlights the compounded effects of multi-phase fairness interventions.

For Statistical Parity Difference (SPD), RisCanvi achieves its lowest value with Re+EGR+EO (0.005), followed closely by DIR+EGR+EO (0.007) and DIR+AL+EO (0.010), all very close to the ideal of zero. Re+AL+ROC (0.024) and Re+EGR+ROC (0.021) show moderate results, while DIR+AL+ROC (0.031) is somewhat higher. COMPAS mirrors this trend: Re+EGR+EO (0.034) and DIR+AL+EO (0.038) achieve the lowest SPD, while DIR+EGR+ROC (0.055) and Re+AL+ROC (0.079) yield higher disparities. Overall, EO-

based integrations consistently deliver stronger parity outcomes than ROC-based ones.

With respect to Disparate Impact (DI), RisCanvi again performs strongly with `Re+EGR+EO` (1.014*), `DIR+EGR+EO` (1.021*), and `DIR+AL+EO` (1.034*), all near the ideal of 1. ROC-based models such as `Re+EGR+ROC` (1.066) and `Re+AL+ROC` (1.092*) diverge further, though still remain moderate. COMPAS shows a similar distribution: EO-based integrations (`Re+EGR+EO`, `DIR+EGR+EO`, `DIR+AL+EO`) cluster close to 1.08–1.09, while ROC-based integrations such as `Re+AL+ROC` (1.159) deviate more. Thus, DI results confirm the advantage of EO-based post-processing in PIP contexts.

For Equal Opportunity Difference (EOD), RisCanvi records some of its lowest disparities across all experiments. `DIR+AL+EO` (0.006), `EGR+EO` (0.033), and `DIR+EGR+EO` (0.011) all approach zero, while ROC-based models such as `Re+AL+ROC` (0.028) and `DIR+EGR+ROC` (0.076) exhibit higher values. COMPAS again shows minimal disparities under EO-based integrations: `Re+EGR+EO` (0.001), `DIR+AL+EO` (0.001), and `DIR+EGR+EO` (0.002) are near-ideal. By contrast, `DIR+AL+ROC` (0.006) remains small, but `Re+AL+ROC` (0.037) is higher. This highlights the robustness of EO-based PIP models in promoting true positive rate equity.

Turning to Predictive Equality Difference (PED), RisCanvi shows very strong outcomes: `DIR+EO`-based combinations such as `DIR+AL+EO` (0.002) and `DIR+EGR+EO` (0.0003) are virtually ideal, while `Re+EGR+EO` (0.010) remains very low. ROC-based results, such as `Re+EGR+ROC` (0.034) and `DIR+EGR+ROC` (0.093), record larger deviations. COMPAS follows a similar trajectory, with EO-based combinations near zero (e.g., `Re+EGR+EO` at 0.0002 and `DIR+AL+EO` at 0.001), while ROC-based ones like `DIR+EGR+ROC` (0.004) and `Re+EGR+ROC` (0.042) are higher. This further emphasizes the superior stability of EO-driven PIP integrations for mitigating false positive rate disparities.

In terms of accuracy, RisCanvi models range from 0.66–0.74. The highest accuracy is achieved by `Re+AL+ROC` (0.74) and `DIR+EGR+ROC` (0.74), though these correspond to weaker fairness results. COMPAS models operate between 0.65–0.68, with

DIR+AL+ROC (0.68) achieving the highest. These outcomes highlight a recurring fairness–accuracy trade-off: ROC-based models often improve accuracy at the expense of fairness, while EO-based models maintain fairness with only modest accuracy reductions.

Comparative Insights. The PIP models illustrate the strengths and limitations of full pipeline integration. EO-based combinations (Re+EGR+EO, DIR+EGR+EO, DIR+AL+EO) deliver near-ideal results across almost all fairness metrics, often outperforming single- or dual-phase integrations, and do so with only moderate reductions in accuracy. In contrast, ROC-based integrations (Re+EGR+ROC, Re+AL+ROC, DIR+EGR+ROC, DIR+AL+ROC) frequently record higher predictive accuracy but at the expense of fairness, amplifying disparities across metrics such as SPD and EOD. These findings suggest that while multi-phase integration enhances flexibility, the choice of post-processing method remains decisive in balancing fairness and accuracy. EO-based PIP models thus emerge as the most promising for contexts where equity considerations are paramount.

5.2.5 Implementation of Many-Objective Optimization

To account for trade-offs between predictive performance and fairness, we implement a Many-Objective Optimization (MOO) approach that simultaneously evaluates five criteria: Accuracy (Acc), Statistical Parity Difference (SPD), Disparate Impact (DI), Equal Opportunity Difference (EOD), and Predictive Equality Difference (PED). Unlike single-metric evaluations, MOO enables a principled comparison across these competing objectives by identifying Pareto-optimal configurations, solutions where no objective can be improved without compromising another.

This framework allows us to formally assess whether fairness interventions outperform baseline models and to identify models that best balance fairness and accuracy. Table 5.1 summarizes the non-dominated solutions across different intervention phases. For the COMPAS dataset, optimal combinations include Re+EGR+EO, DIR+EGR+EO, DIR+AL+ROC, and DIR+AL+EO; for the RisCanvi dataset, DIR+EO and

DIR+AL+EO are identified as optimal. Importantly, the baseline NONE model does not outperform any of the optimal combinations, underscoring the effectiveness of fairness-enhancing strategies in improving both equity and predictive utility.

5.2.6 Bi-Objective Optimization for Targeted Fairness Metrics

In contrast, bi-objective optimization focuses on reconciling two goals, typically one fairness criterion with predictive accuracy. This approach is especially relevant in high-stakes applications, where decision-makers may prioritize specific fairness concerns while maintaining reliable predictive performance.

For the RisCanvi dataset, when optimizing SPD, DI, or EOD jointly with accuracy, models such as AL+ROC, DIR+EO, DIR+ROC, DIR+AL+EO, and Re+AL+ROC consistently appear as non-dominated. When the focus shifts to PED and accuracy, four models dominate the Pareto front: AL+ROC, DIR+EO, DIR+ROC, and Re+AL+ROC.

For the COMPAS dataset, models including DIR+AL, EGR+ROC, and Re+EGR+ROC are Pareto-optimal when SPD or DI is optimized alongside accuracy. When prioritizing EOD and accuracy, the leading models are DIR+EGR+EO, DIR+AL+EO, and DIR+AL+ROC. For PED combined with accuracy, the non-dominated set expands to include Re+AL, DIR+EGR, DIR+EGR+ROC, DIR+AL+EO, and DIR+EGR+EO. These results collectively illustrate that targeted optimization strategies can help tailor fairness–accuracy trade-offs to align with domain-specific requirements.

5.3 Key Insights and Observations

This section reflects on the integration of fairness-enhancing techniques across preprocessing, in-processing, and post-processing stages, with the aim of balancing predictive accuracy and fairness in recidivism prediction. While single-phase fairness interventions demonstrate value, their limitations are evident when compared with multi-stage approaches. The integrated strategies consistently reveal stronger

performance across multiple fairness metrics, as summarized in Table 5.1.

5.3.1 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 empirical evidence indicates that integrated models outperform single techniques in striking a balance between fairness and predictive performance. Several multi-phase combinations achieved considerable reductions in group disparities while sustaining accuracy at competitive levels.

For example, on the RisCanvi dataset, the model `DIR+EGR+EO` attained an SPD of 0.007, DI of 1.021, EOD of 0.011, and PED of 0.0003 with an accuracy of 0.70. This marks a clear advancement compared with the baseline (NONE), which showed higher disparities (SPD = 0.059, DI = 1.224, PED = 0.045) despite slightly better accuracy (0.72). Similarly, `Re+EGR+EO` achieved an even lower SPD (0.005) and DI (1.014) with a marginal accuracy reduction to 0.69.

On the COMPAS dataset, the `DIR+AL+EO` model achieved highly competitive outcomes (SPD = 0.038, DI = 1.089, EOD = 0.001, PED = 0.001) at an accuracy of 0.66. Compared to the baseline model (SPD = 0.25, DI = 2.13, PED = 0.18, Acc = 0.68), this represents significant improvements in all fairness dimensions with only a minor sacrifice in predictive accuracy.

These findings demonstrate that no single fairness technique is sufficient to mitigate bias comprehensively. Integrated approaches leverage complementary strengths across different stages, yielding well-rounded improvements in fairness with limited accuracy trade-offs. This validates the premise that multi-phase fairness pipelines are effective in addressing bias in high-stakes domains.

5.3.2 Impact of Dataset Characteristics on Metrics

RQ2: How does the effectiveness of various fairness metrics vary when applying the integrated approach?

Dataset-specific influences are evident in how fairness metrics behave across RisCanvi and COMPAS. Metrics such as SPD and DI tend to fluctuate more between datasets, reflecting sensitivity to underlying demographic distributions. For example, `Re+EO` achieved DI = 1.035 on RisCanvi but DI = 1.095 on COMPAS, showing heightened dependence on dataset characteristics. In contrast, metrics such as EOD and PED displayed more consistent values across datasets, suggesting greater stability when used in comparative evaluations. These results indicate that fairness assessments should be aligned with dataset properties and objectives, rather than relying on any single metric universally.

5.3.3 Effectiveness of Techniques in Maintaining Accuracy

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

Integrated techniques generally preserve accuracy more effectively than individual ones. On RisCanvi, `DIR+AL+EO` achieved an accuracy of 0.72 while delivering strong fairness (SPD = 0.010, DI = 1.034). On COMPAS, combinations such as `EGR+EO` and `Re+EGR` maintained accuracies of 0.65–0.67 while reducing disparities across metrics. In contrast, individual methods often induced sharper trade-offs. For example, EO alone yielded DI = 1.048 on RisCanvi but lowered accuracy to 0.70. Thus, integrated strategies appear more reliable for balancing fairness with predictive power.

5.3.4 Predictive Stability Across Datasets

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

Integrated approaches improved predictive stability across datasets, ensuring consistent patterns of fairness and accuracy. For example, `DIR+AL+EO` and `Re+EGR+EO` performed comparably on both RisCanvi and COMPAS, with only minor metric variations. By contrast, single-phase methods such as DIR showed variability (SPD = 0.079 on RisCanvi vs. 0.284 on COMPAS). This robustness highlights the advantage of integrated pipelines, which mitigate the compounding effects of bias at multiple stages, thereby increasing reliability in practical applications.

5.3.5 Real-World Applicability and Implications

RQ5: What are the implications of validating fairness-enhanced models on new datasets for real-world applicability?

Evaluating fairness-enhanced models on both RisCanvi and COMPAS highlights their adaptability across diverse contexts. Integrated methods such as `DIR+AL+EO` consistently produced strong fairness outcomes, underscoring their suitability for deployment in sensitive domains like criminal justice. Nevertheless, the variability observed in SPD and DI across datasets shows the importance of tailoring methods to the specific fairness objectives and data characteristics of each application. In practice, this means prioritizing integrated approaches while calibrating them carefully for domain-specific needs.

5.3.6 Why Do Integrated Techniques Behave Differently Across Datasets?

The differences observed across datasets stem from a complex interplay of factors that shape how fairness interventions manifest in practice. These include the intrinsic biases encoded in the datasets, the specific mechanisms by which mitigation techniques operate, the sensitivity of fairness metrics, and the interactions among techniques when combined.

Dataset-Specific Biases. Each dataset reflects unique structural and historical patterns of bias. For example, the COMPAS dataset shows strong correlations be-

tween demographic attributes, particularly race, and recidivism predictions, leading to more pronounced disparities in group fairness metrics such as Statistical Parity Difference (SPD) and Disparate Impact (DI). As a result, pre-processing techniques like the Disparate Impact Remover (DIR) tend to yield more significant improvements on COMPAS. In contrast, the RisCanvi dataset encodes more subtle biases, with weaker associations between demographic features and target outcomes, which limits the effectiveness of such fairness-enhancing techniques.

Moreover, both datasets differ in the degree and nature of class imbalance, which further influences fairness outcomes. Addressing these imbalances is critical for equitable model performance. To this end, this thesis investigates the role of data imbalance in fairness-aware AI and explores oversampling techniques as a potential solution. This is examined in detail in Chapter 6.

Mechanisms of techniques. Fairness interventions act at different stages of the machine learning pipeline, shaping their impact. Pre-processing methods rebalance datasets and thus primarily influence group-level parity measures (SPD, DI). In-processing approaches, such as adversarial learning or Exponentiated Gradient Reduction, enforce fairness constraints during training, yielding stronger effects on error-rate metrics like Equal Opportunity Difference (EOD) and Predictive Equality Difference (PED). Post-processing methods, including Equalized Odds adjustments, operate at the decision stage, fine-tuning classification thresholds to mitigate residual disparities. The effectiveness of these techniques is closely tied to the statistical properties of the dataset in which they are applied.

Metric sensitivity. Fairness metrics themselves exhibit different levels of robustness across datasets. Group disparity measures such as DI and SPD tend to fluctuate significantly with changes in demographic distributions, while error-rate based metrics (EOD and PED) often display more stability across contexts. For example, methods like `EGR+EO` consistently improved EOD and PED on both COMPAS and RisCanvi, demonstrating the relative resilience of these metrics across varying data distributions.

Interactions in integrated models. When multiple fairness techniques are combined, their effects can be complementary or competing, depending on dataset characteristics. For instance, `Re+EGR+EO` displayed strong synergy across both datasets, simultaneously reducing DI, SPD, and EOD while preserving accuracy. This is because Reweighing rebalances group representation, EGR incorporates fairness constraints during learning, and EO adjusts decision thresholds, producing cumulative benefits. On the other hand, `DIR+AL+EO` was highly effective on COMPAS but less so on RisCanvi, where weaker correlations between sensitive attributes and outcomes limited the adversarial component's impact.

Fairness–accuracy trade-offs. Finally, differences in outcomes also reflect the tension between fairness gains and predictive accuracy. Some combinations, such as `Re+EGR+EO`, managed to reduce disparities across all fairness dimensions while incurring only a minor reduction in accuracy. Others, like `DIR+AL+EO`, achieved stronger fairness performance but introduced small accuracy losses, particularly on RisCanvi. These trade-offs underscore the importance of selecting integration strategies that align with the fairness priorities of the application domain.

Taken together, these findings highlight that integrated fairness techniques cannot be assumed to generalize uniformly across datasets. Their effectiveness depends on the underlying biases present in the data, the design of the mitigation methods, the choice of evaluation metrics, and the nature of interactions among combined techniques. This reinforces the need for context-sensitive evaluation when deploying fairness interventions in real-world decision-making systems.

5.3.7 Statistical Significance of Fairness Metrics Across Datasets

This section presents a statistical summary of the impact of fairness-enhancing techniques across two recidivism datasets, RisCanvi and COMPAS, based on key fairness metrics: Statistical Parity Difference (SPD), Disparate Impact (DI), Equal Opportunity Difference (EOD), Predictive Equality Difference (PED), and predictive accuracy (Acc), see Table 5.2. The analysis, which includes mean, standard deviation,

and 95% confidence intervals, reveals stark contrasts in fairness outcomes and consistency between the two datasets, underscoring the context-dependent nature of algorithmic fairness.

Models trained on COMPAS consistently exhibit higher residual bias than those trained on RisCanvi. For instance, the average SPD for COMPAS is 0.0789 compared to 0.0423 on RisCanvi, reflecting a greater disparity in positive outcomes between demographic groups. Similarly, the mean DI on COMPAS is 1.265, deviating substantially from the fairness ideal of 1.0 and indicating that one group receives favorable outcomes about 26% more often than the other. In contrast, RisCanvi achieves a mean DI of 1.154, which, while still above parity, is notably closer to the ideal. This discrepancy extends to other fairness metrics: COMPAS records higher mean EOD (0.0601 vs. 0.0418) and PED (0.0497 vs. 0.0371), reinforcing that its models tend to produce greater intergroup disparities.

Beyond mean differences, the variability of fairness metrics further underscores the instability of fairness outcomes on COMPAS. Standard deviations for COMPAS are roughly double those for RisCanvi across all fairness measures, implying that bias mitigation techniques yield less consistent effects. This pattern is reflected in the width of the 95% confidence intervals. For example, the DI interval for COMPAS [1.1198, 1.4103] lies entirely above 1.0, signifying persistent unfairness across runs, whereas the DI interval for RisCanvi [1.1007, 1.2066], though also above 1.0, indicates a narrower and more stable deviation from parity.

Interestingly, model accuracy remains relatively stable across both datasets. The mean accuracy for RisCanvi (0.7126) slightly exceeds that of COMPAS (0.6648), but both display narrow confidence intervals (e.g., COMPAS CI: [0.6599, 0.6698]) and low variability. This suggests that fairness interventions do not significantly destabilize predictive performance. Thus, while fairness metrics fluctuate considerably across datasets and interventions, accuracy remains a more predictable and robust outcome.

Overall, these findings reaffirm that achieving fairness is considerably more chal-

lenging and unstable on COMPAS, likely reflecting deeper structural and historical biases in that dataset. In contrast, RisCanvi offers a more tractable context, where fairness improvements are both larger in magnitude and more consistent across methods. This highlights the importance of evaluating fairness interventions on multiple datasets: a technique that performs well on a relatively balanced dataset like RisCanvi may not generalize to more historically biased contexts such as COMPAS. Moreover, practitioners must consider not only average fairness levels but also their variability, particularly in high-stakes applications where fairness stability is as critical as predictive performance.

Table 5.2: Summary statistics for key performance and fairness metrics across the RisCanvi and COMPAS datasets. Reported values include the mean, standard deviation, and 95% confidence intervals.

Metric	RisCanvi			COMPAS		
	Mean	Std Dev	95% CI	Mean	Std Dev	95% CI
SPD	0.0423	0.0342	[0.0287, 0.0558]	0.0789	0.0851	[0.0452, 0.1125]
DI	1.1536	0.1338	[1.1007, 1.2066]	1.2650	0.0.3671	[1.1198, 1.4103]
EOD	0.0418	0.0414	[0.0254, 0.0582]	0.0601	0.0882	[0.0252, 0.0950]
PED	0.0371	0.0366	[0.0226, 0.0516]	0.0497	0.0667	[0.0233, 0.0761]
Acc	0.7126	0.0193	[0.7049, 0.7202]	0.6648	0.0125	[0.6599, 0.6698]

5.4 Combining Fairness-Enhancing Approaches: Advantages and Limitations

Based on the comprehensive evaluation conducted in this study, the integration of multiple fairness-enhancing strategies within AI systems reveals both significant benefits and notable challenges. This section outlines key considerations when adopting such combined methods.

5.4.1 Enhanced Fairness Outcomes

Our analysis indicates that leveraging multiple fairness-oriented methods can lead to stronger and more balanced fairness outcomes. This is particularly evident when distinct stages of the AI pipeline are addressed, pre-processing methods mitigate bias in input data, in-processing algorithms target model-level disparities during training, and post-processing techniques fine-tune outcomes to improve fairness at the decision-making stage. The complementary nature of these approaches helps ensure that models better account for group-level disparities, without overly prioritizing one fairness metric at the expense of another.

5.4.2 Greater Flexibility and Model Robustness

Unlike approaches that apply fairness interventions at a single stage, combining methods enables more flexible and adaptable solutions suited for various datasets and application domains. This layered strategy enhances the model's ability to balance predictive accuracy and fairness. For instance, while one method might reduce bias with minimal accuracy loss, another could strengthen predictive performance with limited impact on fairness. This synergy reduces vulnerability to dataset-specific or model-specific biases, resulting in more reliable models across contexts.

5.4.3 Increased Stakeholder Confidence

An integrated fairness strategy can also enhance stakeholder trust, particularly in sensitive domains like criminal justice, where different stakeholders (e.g., policymakers, legal experts, and the public) may value different aspects of fairness. By showing a commitment to fairness across multiple stages of the AI pipeline, such models are more likely to gain broader acceptance and reduce skepticism associated with single-point fairness interventions.

5.4.4 Increased System Complexity

While combining multiple fairness-enhancing strategies lead to improvement, it inevitably introduces added complexity into AI systems. This can make the models more difficult to interpret or audit, potentially obscuring their inner workings. Nevertheless, this complexity may be justified by the improved fairness and trustworthiness achieved. As explainability and transparency in AI remain active research areas, future work should explore how integrated fairness techniques can be designed to remain interpretable and accessible for end-users and auditors.

5.4.5 Higher Computational Requirements

The integration of multiple fairness-enhancing interventions inevitably increases computational overhead, both in terms of processing time and resource consumption. This includes the cumulative cost of applying bias mitigation at several stages of the pipeline and executing many-objective optimization across competing fairness and accuracy metrics. For organizations with limited technical or financial resources, such demands may present a practical barrier to implementation.

While a formal analysis of time or space complexity could yield additional insights into the scalability of these methods, such an investigation was beyond the scope of this study. Given the high variability in hardware environments, software optimizations, and parallelization strategies, runtime comparisons may not generalize meaningfully across contexts. Our primary objective was to empirically evaluate fairness–performance trade-offs across diverse configurations, datasets, and evaluation metrics. Nonetheless, future research could investigate the scalability and efficiency trade-offs of these methods to guide resource-constrained deployments.

Despite these increased costs, we contend that the societal value of fair and equitable decision-making, particularly in high-stakes domains such as criminal justice, justifies the additional computational burden. Ensuring algorithmic fairness must remain a core priority wherever AI systems influence life-altering decisions, including sentencing, release eligibility, and rehabilitation planning.

5.5 Summary

This chapter examined the effectiveness of integrating fairness-enhancing techniques across pre-processing, in-processing, and post-processing stages. The results demonstrated that while single-phase interventions can improve specific fairness metrics, integrated models consistently achieved broader improvements across multiple dimensions of fairness with only minor trade-offs in predictive accuracy. Many-objective and bi-objective optimization analyses further highlighted that no single method is universally effective, but carefully designed combinations, such as `Re+EGR+EO` and `DIR+AL+EO`, emerge as non-dominated solutions on both RisCanvi and COMPAS datasets.

Key insights include: (i) integrated pipelines provide more stable performance across datasets compared to individual methods; (ii) dataset-specific characteristics strongly influence the sensitivity of fairness metrics; and (iii) multi-phase approaches are more effective in balancing fairness and accuracy, enhancing their suitability for real-world deployment in sensitive domains such as criminal justice.

However, these evaluations also reveal that data imbalance, particularly the dominance of majority classes, remains a critical factor shaping both fairness and predictive performance. To explore this further, Chapter 6 investigates oversampling techniques as a strategy to mitigate class imbalance and examines their impact on fairness outcomes in recidivism risk prediction.

While these findings underscore the promise of fairness pipeline integration, they also reveal an important limitation: most evaluations rely on group-level fairness metrics assessed independently by race, gender, or age. Such metrics may obscure harms faced by individuals at the intersection of multiple marginalized identities. To address this, Chapter 7 shifts focus to the concept of intersectional fairness, introducing a framework for auditing compound bias in algorithmic decisions. This transition reflects a broader ethical imperative: achieving fairness not just in aggregate, but also for those most vulnerable to algorithmic harm.

Publication(s) Arising from this Chapter

The work presented in this chapter has been published (or submitted) in the following outlets:

1. Michael Mayowa Farayola, Malika Bendechache, Takfarinas Saber, et al. (2024b). "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

2. Michael Mayowa Farayola, Irina Tal, Takfarinas Saber, et al. (2025). "A fairness-focused approach to recidivism prediction: implications for accuracy, trust, and equity". In: *AI & SOCIETY*, pp. 1–19. DOI: https://doi.org/10.1007/s00146-025-02452-1

These publications reflect the main contributions of this chapter and provide further technical details, extended results, and peer-reviewed validation of the methods and findings.

Chapter 6

Effect of Data Oversampling Techniques on Fairness and Performance

6.1 Introduction

This chapter investigates the impact of oversampling techniques on both classification performance and algorithmic fairness in high-stakes predictive tasks, with a focus on recidivism risk assessment. While oversampling is widely used to correct class imbalance (Carvalho, Pinho, and Brás 2025), its implications for fairness across demographic groups (Rančić, Radovanović, and Delibašić 2021; Sonoda 2023), particularly those defined by sensitive attributes such as race or gender, remain under examined.

Using the COMPAS and RisCanvi datasets, this chapter explored the integration of data oversampling techniques, such as Random Oversampling and various SMOTE-based methods, within machine learning workflows. The evaluation focused on both predictive accuracy and key fairness indicators, including Statistical Parity Difference, Disparate Impact, Equal Opportunity Difference, and Predictive Equality Difference. The findings provide a fairness-aware perspective on how balancing

class distributions can affect both performance and equity outcomes, highlighting the strengths and limitations of these strategies in real-world applications.

6.2 Motivation and Problem Statement

Class imbalance is a persistent issue in real-world datasets, particularly in the criminal justice domain, where majority class dominance can lead to biased learning and disproportionate harms for marginalized groups (Iosifidis and Ntoutsi 2019). In recidivism prediction, for instance, individuals who do not reoffend are often underrepresented, skewing model behavior in ways that may reinforce existing social inequities (Chouldechova 2017).

Oversampling is a common pre-processing technique for mitigating class imbalance by synthetically increasing the representation of minority class instances. Although widely adopted to improve classification performance, its effects on algorithmic fairness, especially in relation to sensitive attributes such as race or gender, have not been thoroughly studied. Much of the existing work prioritizes accuracy, often at the expense of a deeper fairness analysis.

This study aims to address this gap by examining whether oversampling can be used not only as a performance-enhancing strategy but also as a fairness intervention. This question is especially critical in high-stakes contexts such as criminal justice, where algorithmic predictions can significantly affect individuals' lives and liberty.

The central hypothesis of this study posits that fairness-aware oversampling strategies can improve key fairness metrics, such as Disparate Impact, Statistical Parity Difference, Equal Opportunity Difference, and Predictive Equality Difference, while maintaining or even enhancing predictive accuracy. To test this hypothesis, the following research questions guide the investigation:

- **RQ1**: How do fairness-aware oversampling techniques influence recidivism prediction accuracy and fairness across datasets with varying biases?
- **RQ2**: Which fairness-aware oversampling strategy is most effective in address-

ing systemic biases in recidivism datasets?

- **RQ3**: How do different classifiers perform when paired with fairness-aware oversampling methods, and what trade-offs arise between fairness and predictive performance?

- **RQ4**: Can fairness-aware oversampling techniques generalize effectively across recidivism datasets with different levels of bias and class imbalance?

By positioning oversampling as both a technical and ethical lever, this chapter contributes to ongoing discussions on fairness-aware machine learning and the development of responsible AI systems.

6.2.1 Discussion of Dataset Characteristics

This section provides a detailed overview of the characteristics of the COMPAS and RisCanvi recidivism datasets, as introduced in chapter 4. The focus here is on the distributional properties relevant to fairness and performance, including class imbalance, sensitive attributes, applied oversampling methods, algorithm training, and implementation specifics.

Table 6.1 summarizes the distribution of sensitive attributes (s) and outcome labels (y) across the two datasets. The RisCanvi dataset exhibits a more pronounced imbalance, with 1,839 individuals in the $s = 0$ group compared to 1,016 in the $s = 1$ group. The class distribution is similarly skewed, with 2,391 labeled as non-recidivists ($y = 0$) and only 464 labeled as recidivists ($y = 1$). In contrast, the COMPAS dataset contains a larger number of instances overall and shows a comparatively balanced distribution: 3,175 in the $s = 0$ group and 2,103 in $s = 1$, with 2,795 classified as $y = 0$ and 2,483 as $y = 1$.

These disparities underscore the presence of structural bias in both datasets, which can adversely influence model behavior and fairness outcomes. The more severe imbalance in RisCanvi, particularly regarding the under-representation of the positive class ($y = 1$), highlights the necessity of applying oversampling methods.

Such data characteristics present considerable challenges to fairness-aware modeling, reinforcing the need for interventions that address both class and demographic imbalances to promote equitable decision-making in recidivism prediction.

Table 6.1: Distribution of Sensitive Attributes s and Outcome Labels y in the RisCanvi and COMPAS Datasets

RisCanvi				**COMPAS**			
s = 0	**s = 1**	**y = 0**	**y = 1**	**s = 0**	**s = 1**	**y = 0**	**y = 1**
1839	1016	2391	464	3175	2103	2795	2483

6.2.2 Oversampling Approaches

This study adopts four distinct oversampling strategies to address various forms of imbalance in the data. The first approach applies traditional oversampling techniques, such as RandomOverSampler and SMOTE, to correct class imbalance between positive ($y = 1$) and negative ($y = 0$) outcomes, without regard to demographic group membership. The second strategy focuses on demographic representation by oversampling only from the underrepresented (discriminated) group ($s = 0$), independent of the class label. The third approach targets intra-group imbalance by equalizing the number of desired ($y = 1$) and undesired ($y = 0$) instances within the discriminated group. Finally, the fourth strategy ensures balanced distributions of outcomes ($y = 1$ and $y = 0$) within both the privileged and discriminated groups separately, before recombining them into a unified, balanced dataset for training.

To operationalize these strategies, we implemented a suite of eight oversampling methods, summarized in Table 6.2, including both basic and hybrid techniques. RandomOverSampler (ROS) (Estabrooks, Jo, and Japkowicz 2004) serves as a baseline by replicating instances from the minority class. Several SMOTE-based techniques were also employed, namely, SMOTE (Chawla et al. 2002), Borderline-SMOTE (H. Han, W.-Y. Wang, and Mao 2005), KMeans-SMOTE (Douzas, Bacao, and Last 2018), SMOTEN (Ratih et al. 2022), SVM-SMOTE (J.-B. Wang, C.-A. Zou, and Fu 2021), SMOTE-Tomek (Batista, Bazzan, Monard, et al. 2003), and SMOTEENN

Table 6.2: Oversampling Techniques and Their Implementation in Python

No.	Oversampling Technique	Python Implementation
1	Random OverSampler	`RandomOverSampler()`
2	Synthetic Minority Over-sampling Technique (SMOTE)	`SMOTE()`
3	SMOTE + Tomek Links (SMOTETomek)	`SMOTETomek()`
4	SMOTE with Edited Nearest Neighbors (SMOTEENN)	`SMOTEENN()`
5	SMOTE for Nominal Features (SMOTEN)	`SMOTEN(k_neighbors=5)`
6	KMeans Clustering with SMOTE (KMeans-SMOTE)	`KMeansSMOTE(k_neighbors=5)`
7	Borderline SMOTE (Bo-SMOTE)	`BorderlineSMOTE()`
8	SVM-Based SMOTE (SVMSMOTE)	`SVMSMOTE()`

(Xu et al. 2020), each generating synthetic samples to better represent minority class regions and decision boundaries (Chawla et al. 2002; H. Han, W.-Y. Wang, and Mao 2005; Batista, Prati, and Monard 2004). The hybrid variants, SMOTE-Tomek and SMOTEENN, combine oversampling with undersampling methods (Tomek links and edited nearest neighbors, respectively) to both balance the data and eliminate noisy or borderline instances.

These oversampling techniques were selected for their documented effectiveness in handling imbalanced datasets and their relevance in fairness-aware machine learning research (Carvalho, Pinho, and Brás 2025; Mohammed, Rawashdeh, and Abdullah 2020; Kabir et al. 2024). Beyond improving classifier performance, this study investigates how these techniques influence fairness outcomes across sensitive demographic groups, providing a comprehensive analysis of fairness-accuracy trade-offs in the context of recidivism prediction.

6.2.3 Model Training, Evaluation and Procedure

This study employed seven classification algorithms, Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, Support Vector Machine, k-Nearest Neighbors, and Gaussian Naive Bayes, to assess the effectiveness of various oversampling methods on fairness and predictive accuracy. Each classifier was paired with oversampling techniques using pipelines that incorporated three-fold cross-validation and hyperparameter optimization via grid search. Following model training on the

resampled training data, performance and fairness were evaluated using a separate test set to ensure unbiased assessment.

The evaluation involved both conventional performance metrics and fairness-specific indicators. Accuracy, defined as the proportion of correctly predicted instances, served as the primary performance metric. Fairness was assessed using four widely adopted measures. *Disparate Impact (DI)* evaluates the ratio of favorable outcomes between unprivileged and privileged groups, with a value of 1 signifying equitable treatment. *Statistical Parity Difference (SPD)* quantifies differences in favorable outcome rates, with a value of 0 indicating parity. *Equal Opportunity Difference (EOD)* measures disparities in true positive rates (TPR), and *Predictive Equality Difference (PED)* captures variations in false positive rates (FPR); both aim for a value of 0 to represent fairness.

These metrics were chosen for their complementary focus on fairness from both allocation and error-based perspectives. DI and SPD examine whether benefits are distributed equitably regardless of true labels, while EOD and PED consider the fairness of classification errors. This combination of metrics is particularly relevant in the criminal justice domain, where imbalanced treatment, especially in terms of prediction errors, can carry serious societal consequences.

In each experiment, oversampling was applied exclusively to the training data using one of the defined techniques. The classifier was subsequently trained, hyperparameters optimized, and predictions made on the test set. Final evaluations were based on both predictive performance and fairness outcomes.

6.2.4 Implementation Details

All experiments were conducted using Python. The implementation relied on several key libraries: `scikit-learn` for model training and evaluation, `imbalanced-learn` for oversampling methods, and `AIF360` for fairness auditing. To guarantee reproducibility across all experimental conditions, a fixed random seed (`random_state=42`) was used consistently throughout the workflow. This consistent setup ensured a re-

Table 6.3: Classifiers and Corresponding Parameter Grids for Tuning

Name	Model	Parameter 1	Parameter 2
LR	Logistic Regression	`classifier__C: [0.1, 1, 10]`	`classifier__solver: ['liblinear', 'lbfgs']`
KNN	k-Nearest Neighbors	`classifier__n_neighbors: [3, 5, 7]`	`classifier__weights: ['uniform', 'distance']`
RF	Random Forest	`classifier__n_estimators: [50, 100]`	`classifier__max_depth: [None, 10]` `classifier__min_samples_split: [2, 5]`
SVM	Support Vector Machine	`classifier__C: [0.1, 1, 10]`	`classifier__kernel: ['linear', 'rbf']`
DT	Decision Tree	`classifier__max_depth: [None, 10, 20]`	`classifier__min_samples_split: [2, 5, 10]`
NB	Gaussian Naive Bayes	*None*	
GraB	Gradient Boosting	`classifier__n_estimators: [50, 100]`	`classifier__learning_rate: [0.1, 0.5]` `classifier__max_depth: [3, 5]`

liable foundation for exploring the impact of different oversampling strategies on model behavior and fairness.

6.3 Comparative Analysis of Oversampling Techniques and Strategies

This section provides a comprehensive comparative evaluation of multiple oversampling strategies applied to the COMPAS and RisCanvi recidivism datasets. Each combination of oversampling technique and classification model was assessed in terms of both accuracy and fairness metrics. The analysis identifies which oversampling strategies are most effective at reducing group-based disparities while preserving or enhancing model performance. The results contribute to a deeper understanding of the trade-offs inherent in fairness-aware data preprocessing, particularly within the context of high-stakes decision-making systems such as criminal justice.

6.3.1 Traditional Oversampling

In the COMPAS dataset, traditional oversampling methods such as RandomOverSampler, KMeans-SMOTE, and SMOTETomek demonstrated moderate improvements across various fairness measures (see Figure 6.1). For instance, when used

with the k-Nearest Neighbors (KNN) classifier, RandomOverSampler achieved a Disparate Impact (DI) of 1.53 and a Statistical Parity Difference (SPD) of 0.15, indicating a notable reduction in group disparities. KMeans-SMOTE combined with KNN attained the lowest Equal Opportunity Difference (EOD) of 0.14, whereas SMOTETomek paired with Naive Bayes (NB) achieved the best Predictive Equality Difference (PED) of 0.08. However, methods like SMOTEENN occasionally resulted in DI values exceeding 2.0, suggesting the possibility of overcompensation. Similarly, Borderline-SMOTE with a Decision Tree (DT) yielded a DI of 2.19, underscoring how overly aggressive sampling can distort fairness metrics if not carefully managed. Across models, KMeans-SMOTE and RandomOverSampler provided the most consistent results in terms of DI and SPD, especially under KNN and Logistic Regression (LR).

Regarding predictive performance (Figure 6.2), the best overall outcome was obtained by combining Borderline-SMOTE with Gradient Boosting (GraB), which achieved an F1-Score of 0.6448, Recall of 0.6408, and AUC-ROC of 0.7201. GraB demonstrated superior performance across multiple oversampling strategies (F1-Score range: 0.6297–0.6448). In contrast, KMeans-SMOTE paired with NB resulted in the lowest F1-Score of 0.4305, despite producing a decent AUC-ROC of 0.6926, suggesting poor classification of positive cases. SMOTEENN with DT recorded a high F1-Score of 0.5984 but a relatively low AUC-ROC of 0.6631, indicative of overfitting. Notably, RandomOverSampler used with Support Vector Machine (SVM) delivered the highest Recall (0.6231) but only moderate F1-Score (0.6366), illustrating the importance of evaluating both sensitivity and precision.

In the RisCanvi dataset (Figure 6.3), KMeans-SMOTE and SMOTEENN were the most effective in improving fairness outcomes. SMOTEENN combined with KNN resulted in a DI of 0.86 and EOD of -0.02. Meanwhile, KMeans-SMOTE used with SVM produced the most balanced SPD at -0.007, and the best PED value of -0.001 was obtained when KMeans-SMOTE was paired with GraB. However, LR combined with SMOTETomek led to a DI of 0.36 and SPD of -0.23, signal-

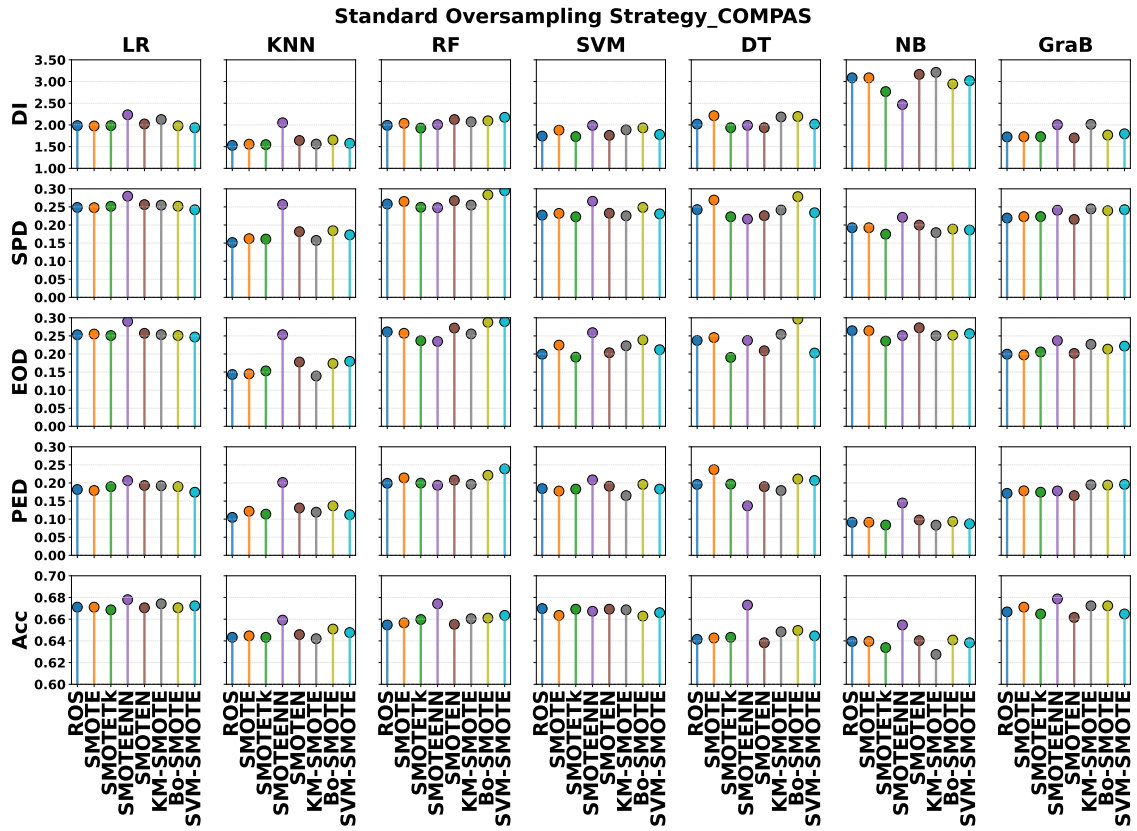

Figure 6.1: Standard Oversampling Strategy - COMPAS

ing insufficient mitigation of group disparities. In terms of classification accuracy, KMeans-SMOTE with Random Forest (RF) achieved the highest at 84%.

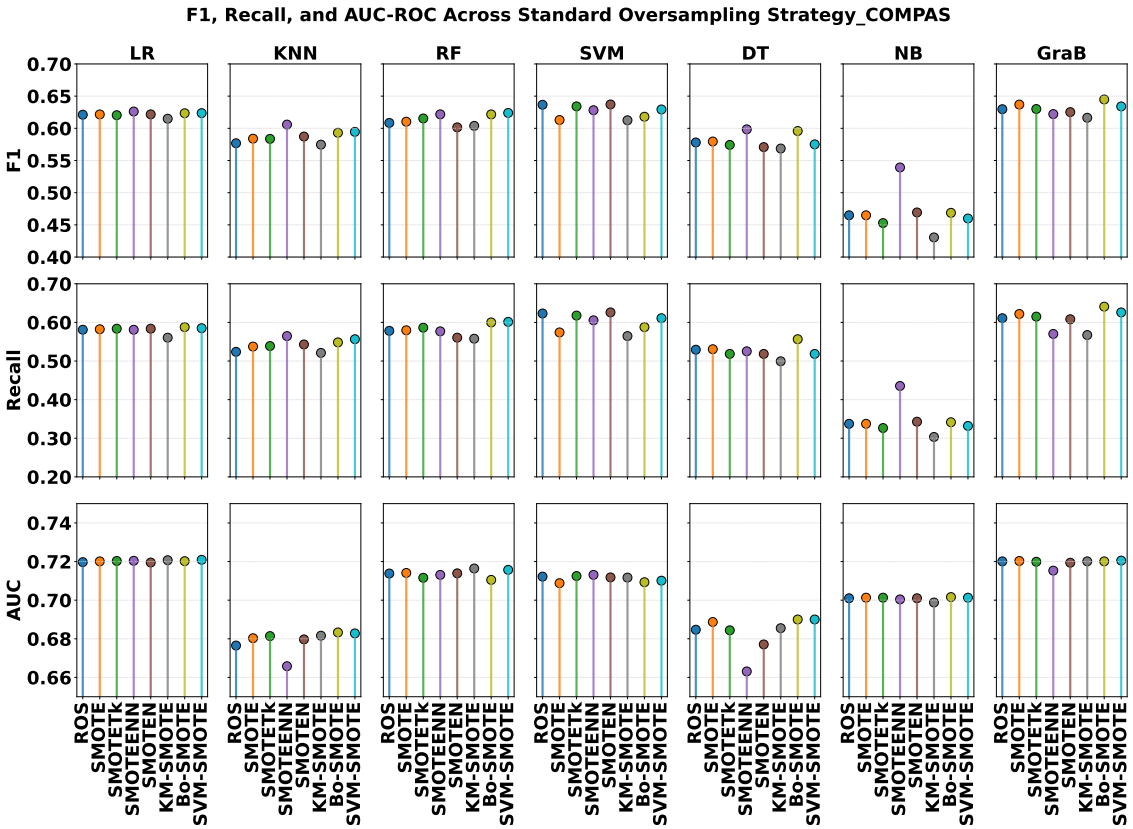

Figure 6.2: Standard Oversampling Strategy - COMPAS

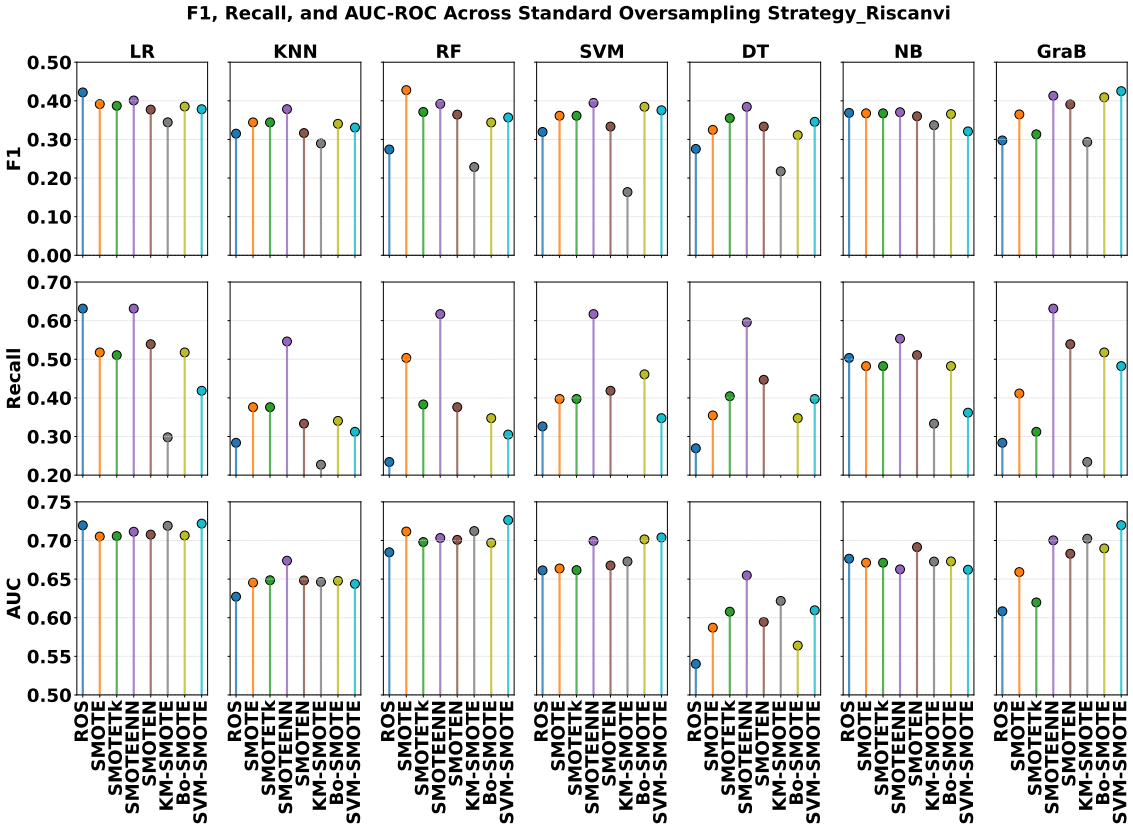

Figure 6.4: Standard Oversampling Strategy - RisCanvi

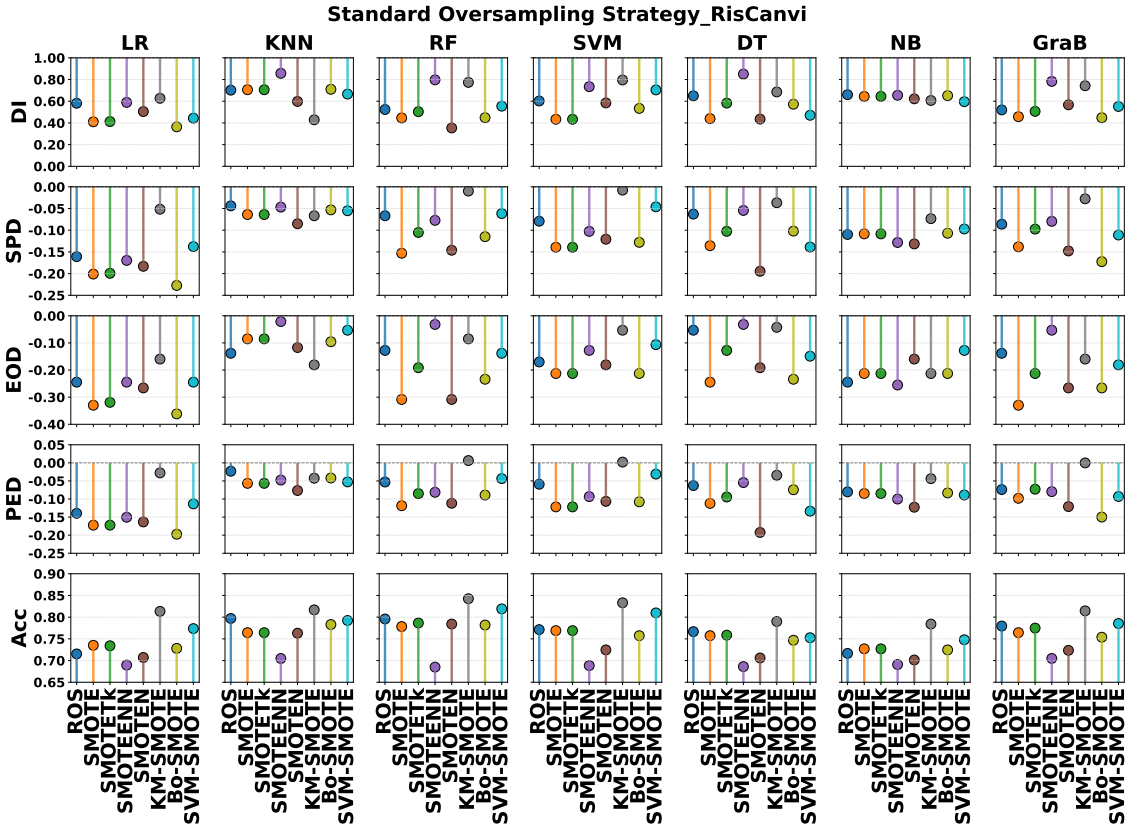

Figure 6.3: Standard Oversampling Strategy - RisCanvi

When analyzing predictive metrics (Figure 6.4), SMOTEENN with LR showed the strongest results, reaching an F1-Score of 0.4009 and Recall of 0.6312. LR and GraB consistently provided reliable outcomes (F1 $>$ 0.39 across multiple oversampling methods). Conversely, SVM paired with SMOTETomek performed poorly (F1: 0.0139, Recall: 0.0071), revealing a mismatch between cleaning-based oversampling and margin-sensitive models like SVM. Another illustrative trade-off was seen with RandomOverSampler and NB, which yielded a relatively high Recall (0.5035) but a low F1-Score (0.3688), indicating a prevalence of false positives. Interestingly, SMOTEENN with NB attained a high AUC-ROC of 0.7315 despite a low F1-Score, highlighting the potential benefit of post-training threshold adjustment to enhance predictive stability.

Insight: While traditional oversampling methods can provide moderate fairness improvements, they may not fully resolve entrenched dataset biases. KMeans-SMOTE paired with KNN or LR generally offered a balanced compromise, achieving

moderate fairness gains without compromising performance. However, techniques like SMOTEENN and SMOTETomek introduced notable variance, with DI values fluctuating below 0.5 or above 2.0, signaling risks of under- or over-adjustment. From a performance standpoint, GraB and LR models, particularly when combined with Borderline-SMOTE or SMOTEENN, delivered the most balanced trade-offs. The inclusion of F1, Recall, and AUC-ROC alongside fairness metrics reveals the need for post-training calibration to optimize reliability, especially in imbalanced datasets like RisCanvi. Overall, GraB combined with Borderline-SMOTE appears to be a promising starting point for practitioners focused on balancing fairness and predictive strength, though care must be taken to avoid metric instability in real-world deployment.

6.3.2 Oversampling Based on Sensitive Attributes

Applying sensitive attribute-based oversampling to the COMPAS dataset led to modest improvements in fairness outcomes relative to traditional strategies (see Figure 6.5). Among the techniques evaluated, SMOTEENN, when combined with Decision Trees (DT), yielded a Disparate Impact (DI) score of 1.05 and performed well across other fairness metrics, achieving an SPD of 0.05, EOD of 0.06, and PED of -0.01 when paired with Gradient Boosting (GraB). Logistic Regression (LR), in conjunction with RandomOverSampler, attained an overall accuracy of 68%.

Additional fairness assessments revealed that the pairing of Borderline-SMOTE with DT maintained relatively stable group parity, with SPD and EOD values of -0.0218 and -0.0426, respectively. GraB demonstrated a consistent capacity to minimize false positive disparities, achieving PED values near zero (e.g., -0.0036) when used with SMOTE or SMOTEENN. However, elevated classification accuracy did not always correspond to equitable outcomes. For instance, although GraB attained a high accuracy of 83.08%, it showed a low DI score of 0.41, implying potential fairness sacrifices despite strong performance.

In terms of performance (see Figure 6.6), the combination of RandomOverSam-

pler and LR produced the highest F1-score (0.6186) and AUC-ROC (0.7199), confirming LR's effectiveness when paired with simple oversampling strategies. Overall, both LR and Random Forest (RF) maintained robust performance across multiple methods, with AUC-ROC consistently above 0.71. In contrast, SMOTEENN tended to degrade classifier performance, particularly for Naive Bayes (NB) and KNN, where AUC-ROC dropped to as low as 0.6625. On a more positive note, SVM-SMOTE in conjunction with RF delivered a balanced profile (F1: 0.6147, Recall: 0.5741), suggesting that more nuanced synthetic sampling could benefit complex models handling demographic imbalance.

In the RisCanvi dataset (see Figure 6.7), sensitive attribute-based oversampling yielded diminished results due to the dataset's skewed group distribution, particularly the disproportion between $s = 1, y = 0$ and $s = 1, y = 1$. This imbalance biased models toward majority predictions and limited the effectiveness of fairness interventions. DI scores across methods remained below the desired threshold (e.g., 0.45–0.51) when using SMOTETomek and SMOTE with classifiers such as GraB, LR, and KNN. Even in cases where SPD and PED were close to zero, such as with Borderline-SMOTE and NB (SPD = -0.019, PED = -0.047), DI and EOD metrics still reflected notable group disparities.

Performance-wise (see Figure 6.8), RandomOverSampler with NB produced the highest F1-score (0.3740) and Recall (0.5106), highlighting NB's competence under this oversampling strategy. NB also maintained Recall above 0.48 across various configurations, although fairness metrics remained inconsistent. LR and RF displayed relatively strong AUC-ROC values (some exceeding 0.72), albeit with more modest F1-scores, pointing to the potential value of threshold adjustment post-training.

Notably, SVM coupled with SMOTEENN resulted in zero Recall and F1, indicating a failure to generalize from oversampled data. More broadly, SMOTE frequently impaired SVM's effectiveness (F1 as low as 0.0139), whereas GraB, when paired with SMOTEENN, offered a more stable yet modest performance (F1: 0.2636, AUC-ROC: 0.6643).

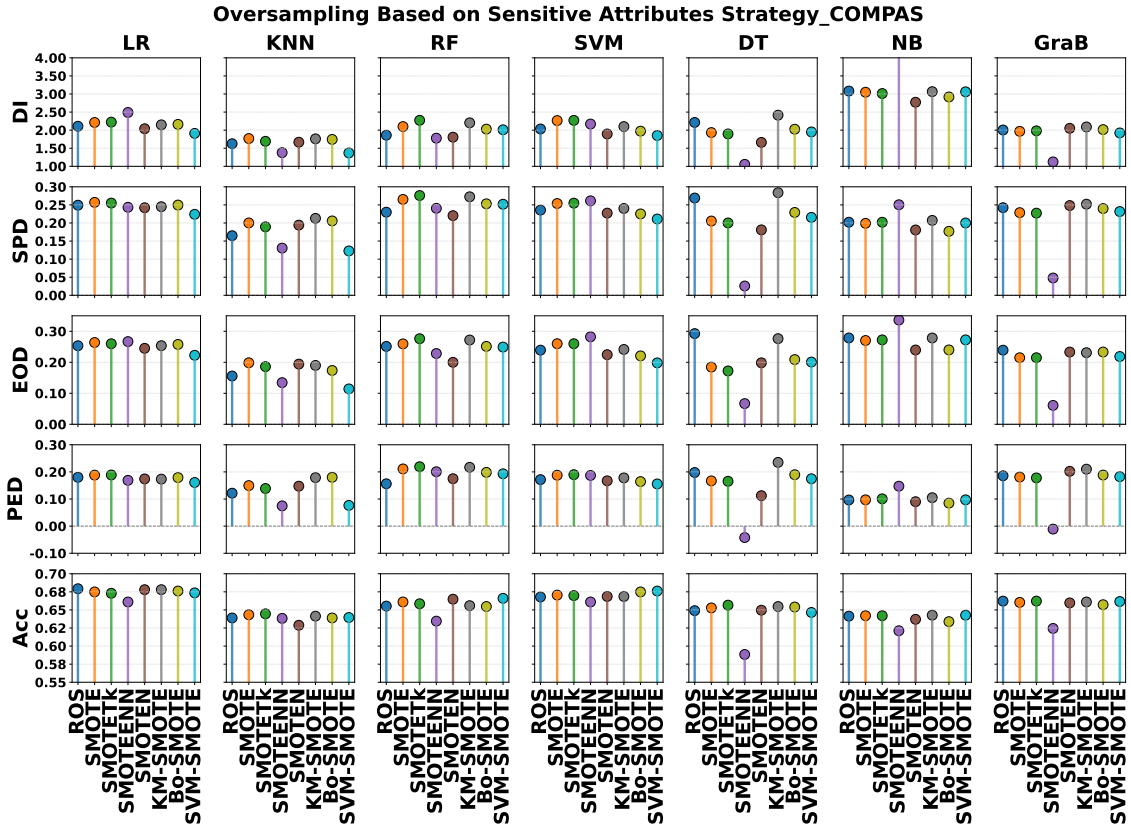

Figure 6.5: Sensitive Attribute Oversampling - COMPAS

Insight: Sensitive attribute-based oversampling achieved some fairness improvements in the COMPAS dataset but proved less effective in the more imbalanced RisCanvi dataset. While GraB and LR demonstrated strengths in minimizing error-based disparities such as PED and EOD, these gains did not always extend to distributional fairness metrics like DI. Simpler approaches like RandomOverSampler provided consistent results with LR and NB, whereas complex methods such as SMOTEENN introduced volatility, especially when used with classifiers like SVM. These results emphasize the importance of selecting oversampling strategies that align with classifier behavior and dataset characteristics in fairness-critical applications.

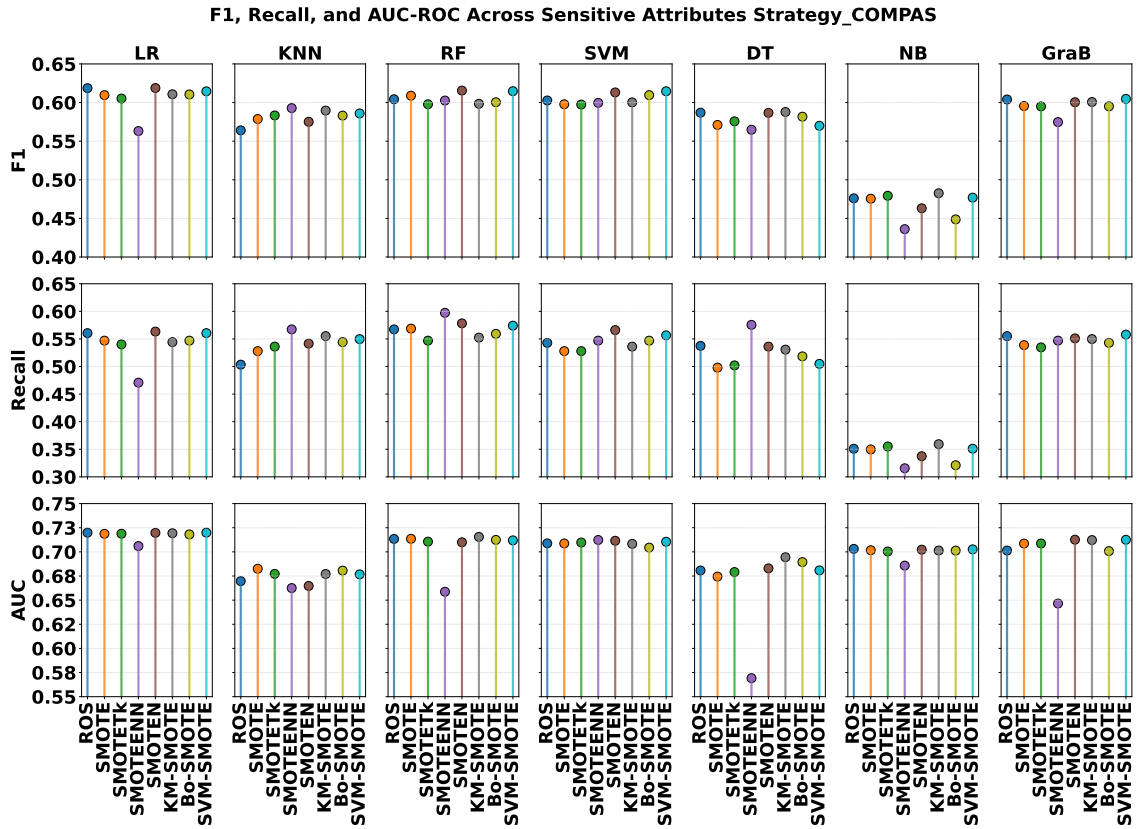

Figure 6.6: Sensitive Attribute Oversampling - COMPAS

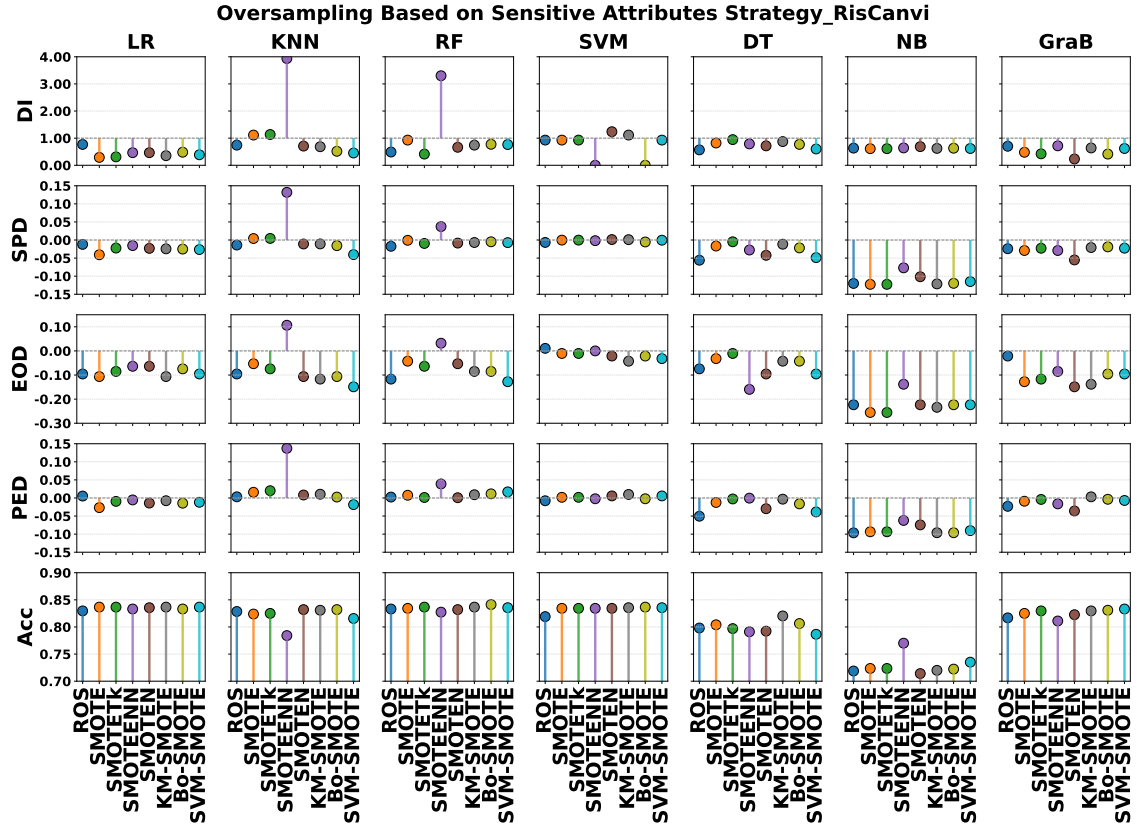

Figure 6.7: Sensitive Attribute Oversampling - RisCanvi

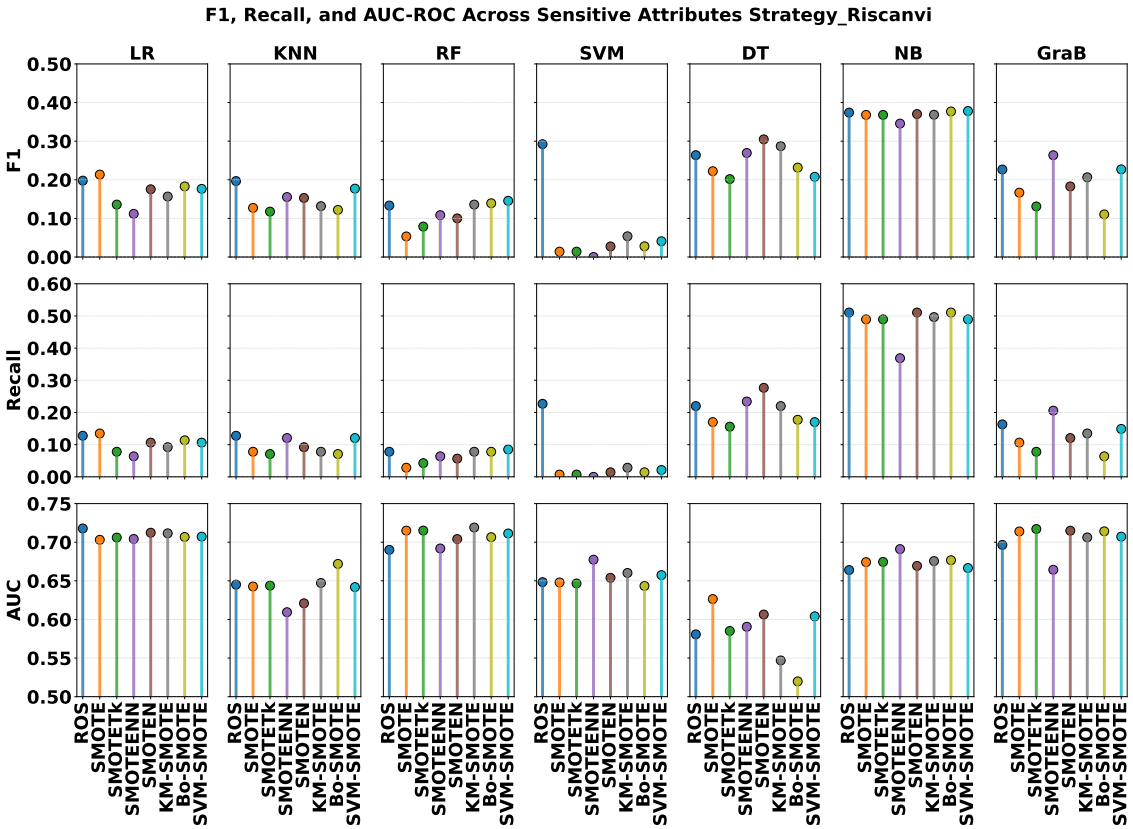

Figure 6.8: Sensitive Attribute Oversampling - RisCanvi

6.3.3 Equalized Discriminated Group Instances

The application of this strategy to the COMPAS dataset (Figure 6.9) produced mixed fairness outcomes. Among all configurations, SMOTEN combined with the k-Nearest Neighbors (KNN) classifier performed best in terms of group fairness, yielding a Disparate Impact (DI) of 1.31, Statistical Parity Difference (SPD) of 0.11, and Equal Opportunity Difference (EOD) of 0.07. For Predictive Equality Difference (PED), the highest value of 0.07 was observed when Bo-SMOTE was paired with Naive Bayes (NB). Predictive accuracy remained relatively stable across models, with SMOTEN and Logistic Regression (LR) achieving the highest accuracy at 68%.

Broader fairness evaluation revealed certain trends. Although SMOTEN with KNN showed promising parity, other combinations introduced instability. For instance, Borderline-SMOTE with Decision Tree (DT) reached a DI of 1.93, while both Gradient Boosting (GraB) and LR also exceeded DI = 1.90, suggesting poten-

tial over-adjustment. EOD and SPD metrics peaked around 0.20 for GraB and LR but remained more moderate for KNN (EOD = 0.0859, SPD = 0.1148). Meanwhile, the lowest PED was recorded with NB (0.0701), pointing to its ability to reduce disparity in false positive rates.

In terms of predictive performance (Figure 6.10), the best F1-Score (0.6302) and Recall (0.6272) came from the combination of SMOTEENN and Random Forest (RF), though its AUC-ROC was slightly lower (0.6919) compared to other top-performing configurations. A balanced profile was also seen with Borderline-SMOTE and GraB, which yielded a slightly lower F1 (0.5997) but a higher AUC-ROC (0.7187), indicating stronger confidence in prediction ranking. Both RF and GraB consistently produced F1-Scores above 0.60 when paired with SMOTEENN or Bo-SMOTE. On the other hand, SMOTETomek with DT showed the weakest performance, achieving an F1 of 0.5234 and AUC-ROC of 0.6514, reflecting poor generalization.

In the RisCanvi dataset (Figure 6.11), fairness values frequently exceeded desirable thresholds, especially with SMOTEENN. For example, LR and DT reached extreme DI values of 2.70 and 2.43 when paired with SMOTEENN and Borderline-SMOTE, respectively, indicating significant overcorrection. More balanced fairness outcomes were achieved using SMOTEN with NB, which recorded DI = 0.8868, SPD = -0.0272, and PED = -0.0043. Bo-SMOTE combined with KNN achieved an EOD of 0.0000, signaling parity in true positive rates. GraB consistently maintained balanced metrics, reporting DI = 1.86, SPD = 0.0769, and PED = 0.0735.

Although predictive accuracy was slightly lower in some cases, most values ranged between 72% and 83%, with KMeans-SMOTE paired with RF and SVM achieving the highest accuracy at 83%. GraB, LR, and KNN also showed strong performance, exceeding 79%. However, fairness was not always in line with accuracy. For instance, LR achieved 79.93% accuracy but reported a DI of 2.7043, indicating a fairness-accuracy trade-off due to aggressive oversampling.

From a performance lens (Figure 6.12), LR paired with SMOTEENN reached

the highest F1-Score (0.3368) and Recall (0.3404), although its AUC-ROC was relatively modest (0.6635). The top-performing F1-Score overall was observed with NB and Borderline-SMOTE (F1 = 0.3717). Meanwhile, SMOTEENN with DT offered higher Recall (0.2979) but posted the lowest AUC-ROC (0.5766), indicating poor confidence in prediction ranking. The weakest configuration was RF with KMeans-SMOTE, which recorded an F1 of 0.1538 and Recall of 0.0922, suggesting that clustering-based oversampling is ill-suited for datasets with limited minority representation.

Insight: The equalized discriminated group instance strategy shows promise in addressing group-level disparities but is sensitive to oversampling aggressiveness. Methods like SMOTEENN can result in inflated fairness metrics that may mask issues in ranking confidence (AUC-ROC) or model generalizability. For the COMPAS dataset, RF and GraB maintained the most consistent trade-off between fairness and performance, whereas for RisCanvi, simpler classifiers like LR and NB handled oversampling more reliably. These outcomes emphasize the importance of aligning resampling methods with both model complexity and data structure. Moreover, discrepancies between Recall and AUC-ROC suggest a need for post-training calibration or threshold optimization to ensure fairness interventions do not compromise real-world applicability.

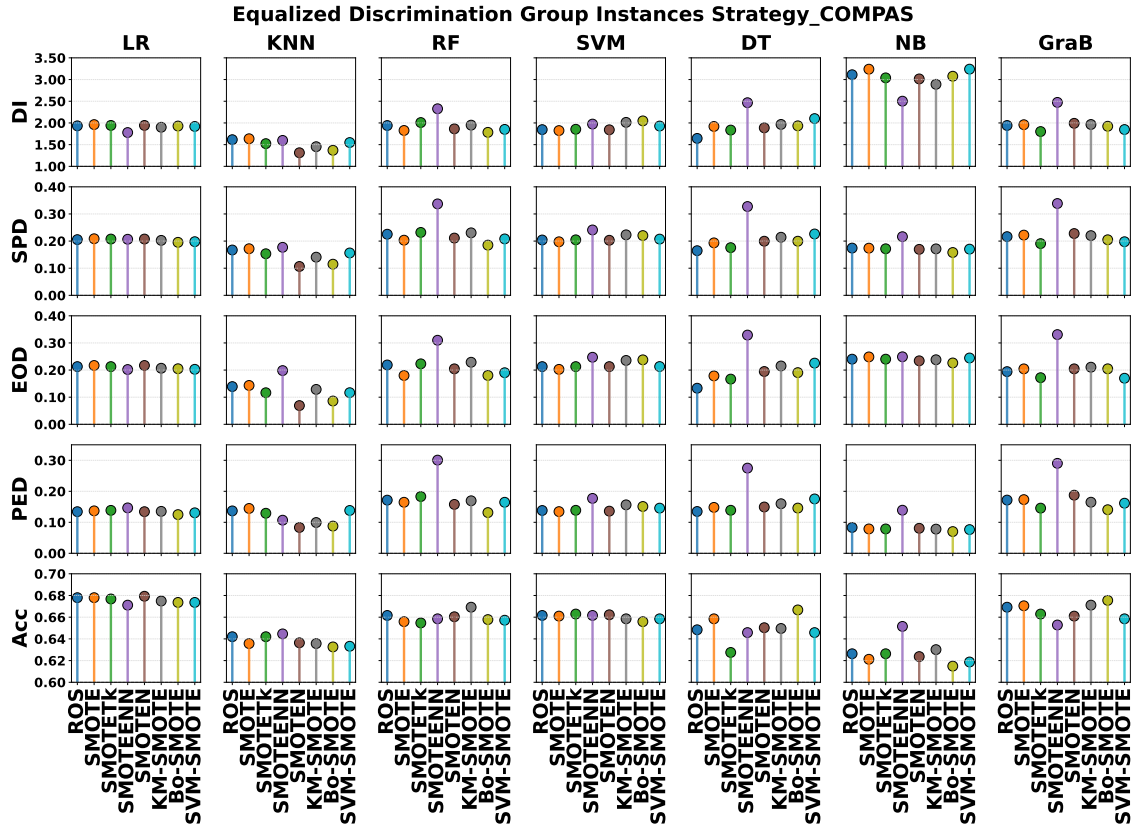

Figure 6.9: Equalized Discrimination Strategy - COMPAS

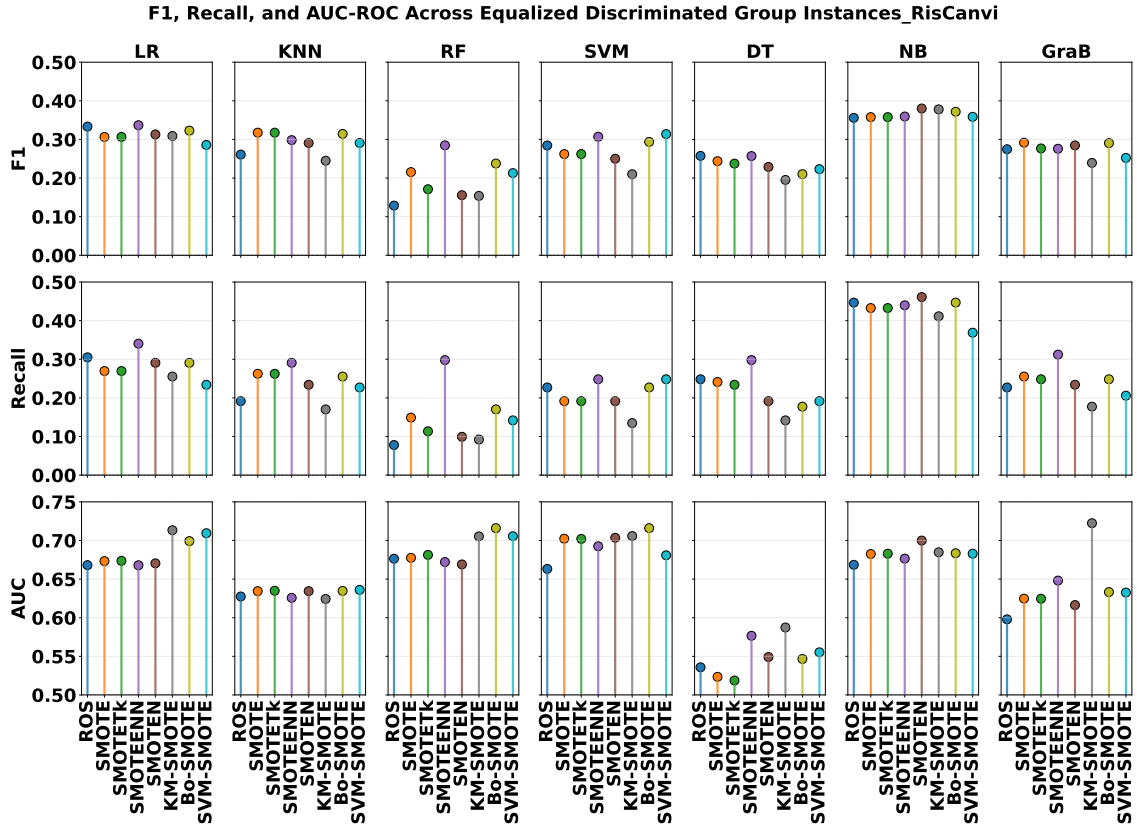

Figure 6.12: Equalized Discrimination Strategy - RisCanvi

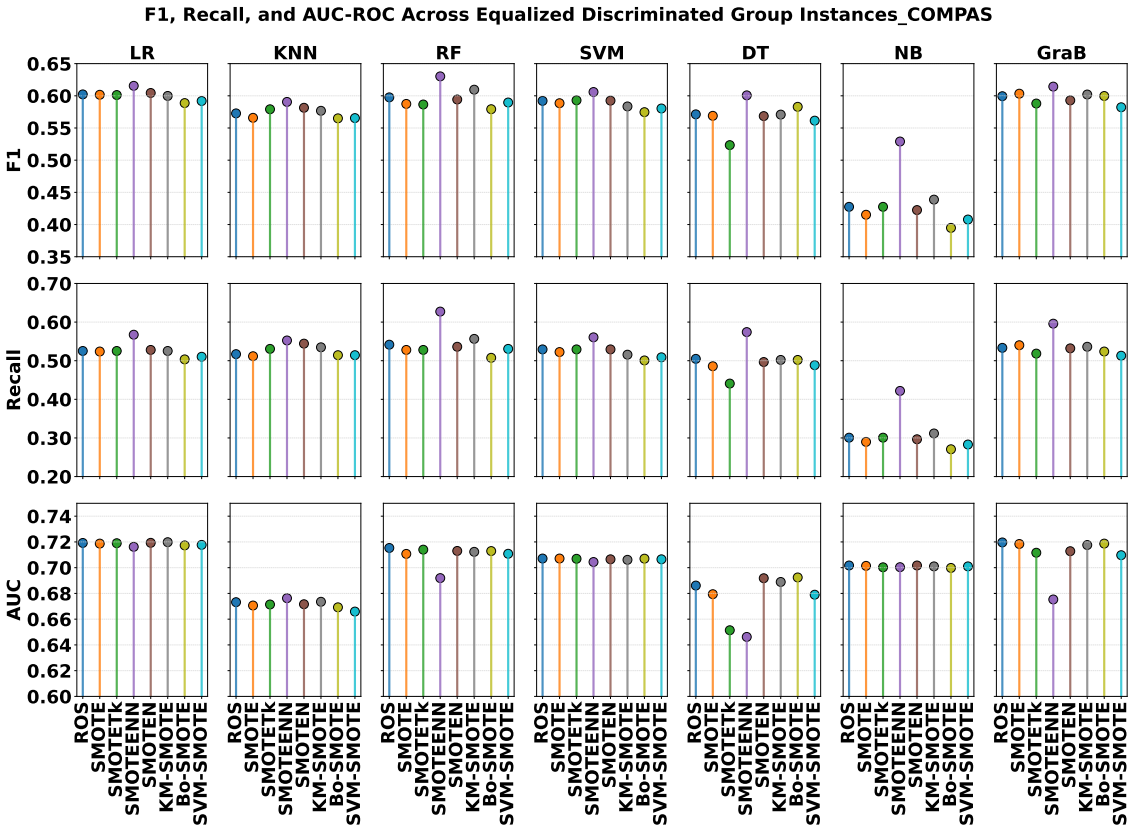

Figure 6.10: Equalized Discrimination Strategy - COMPAS

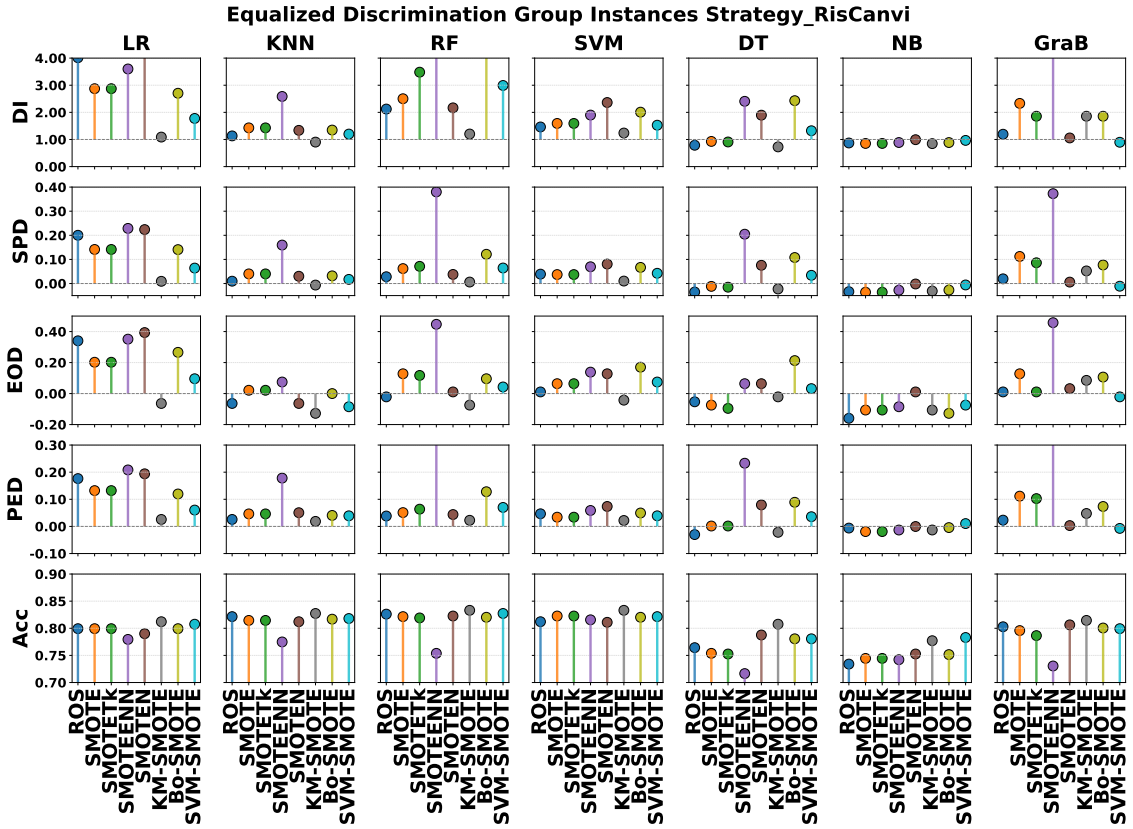

Figure 6.11: Equalized Discrimination Strategy - RisCanvi

6.3.4 Equalized Desired Outcomes Across Groups

This strategy produced the most consistently favorable fairness results in the COMPAS dataset, as shown in Figure 6.13. Among all classifier-oversampler combinations, Borderline-SMOTE combined with Gradient Boosting (GraB) yielded near-perfect group fairness, with a Disparate Impact (DI) of 0.9949 and a Statistical Parity Difference (SPD) of -0.0022. Additionally, RandomOverSampler paired with GraB delivered the best Equal Opportunity Difference (EOD) at -0.0028, while SMOTEN with GraB achieved the highest Predictive Equality Difference (PED) of 0.0048. Across different configurations, fairness metrics consistently hovered around ideal thresholds (DI $\approx$ 1, SPD $\approx$ 0), demonstrating the strategy's capability to achieve near-parity.

In terms of predictive performance (Figure 6.14), SMOTEN combined with Support Vector Machine (SVM) recorded the highest F1-Score of 0.6340 and Recall of 0.6259, indicating a strong balance between precision and sensitivity. Both SVM and Logistic Regression (LR) showed stable performance, consistently achieving F1-Scores above 0.61 when used with SMOTEN or SVM-SMOTE. While SMOTEENN and SVM delivered a Recall of 0.6190, its slightly lower F1-Score (0.6203) reflected a typical trade-off between precision and recall. KMeans-SMOTE was generally less competitive; however, when paired with LR, it reached an AUC-ROC of 0.7180, demonstrating LR's resilience in handling synthetic data. In contrast, Naive Bayes (NB) underperformed, particularly when combined with Borderline-SMOTE (F1: 0.4067, Recall: 0.2816), highlighting its limitations under complex oversampling conditions.

Fairness results were even more consistent in the RisCanvi dataset (Figure 6.15). The pairing of RandomOverSampler with Decision Tree (DT) led to optimal fairness, with a DI of 1.0056 and SPD of 0.0011. SMOTEENN with LR achieved the most equitable EOD at -0.0032, while SMOTEN with GraB attained the lowest PED value of -0.001, indicating minimal disparity in false positive rates. Notably, GraB consistently maintained fairness across all indicators, posting DI (0.8900), SPD (-

0.0246), and PED (-0.0011), all well within acceptable thresholds.

In terms of classification outcomes (Figure 6.16), the combination of SVM-SMOTE and LR yielded the highest F1-Score (0.4422) alongside a strong AUC-ROC of 0.7596. LR and Random Forest (RF) also demonstrated robust adaptability, frequently exceeding F1-Scores of 0.40 with either SVM-SMOTE or SMOTEENN. For high-sensitivity predictions, SMOTEENN with RF achieved the highest Recall (0.6691), although its F1-Score was slightly lower (0.4043), suggesting a trade-off in precision. The best AUC-ROC performance was achieved by GraB with KMeans-SMOTE (0.7646), showing strong ranking capabilities despite more modest Recall and F1 values. On the other hand, the combination of DT and KMeans-SMOTE delivered the weakest results (F1: 0.3009, AUC-ROC: 0.5918), underscoring DT's limitations in noisy data environments.

Insight: The Equalized Desired Outcomes Across Groups approach demonstrated superior performance in balancing fairness and accuracy across both datasets. In COMPAS, GraB and LR consistently promoted equitable treatment, while SVM paired with SMOTEN excelled in classification. For RisCanvi, LR and RF combined with fairness-aware oversamplers like SVM-SMOTE and SMOTEENN yielded strong results across both fairness and performance metrics. The consistently high AUC-ROC values suggest that these models produce reliable risk scores, although observed disparities between AUC and F1 metrics reinforce the need for threshold calibration to optimize decision-making. Overall, this strategy emerged as the most effective in achieving fair and accurate outcomes when properly matched with model and data characteristics.

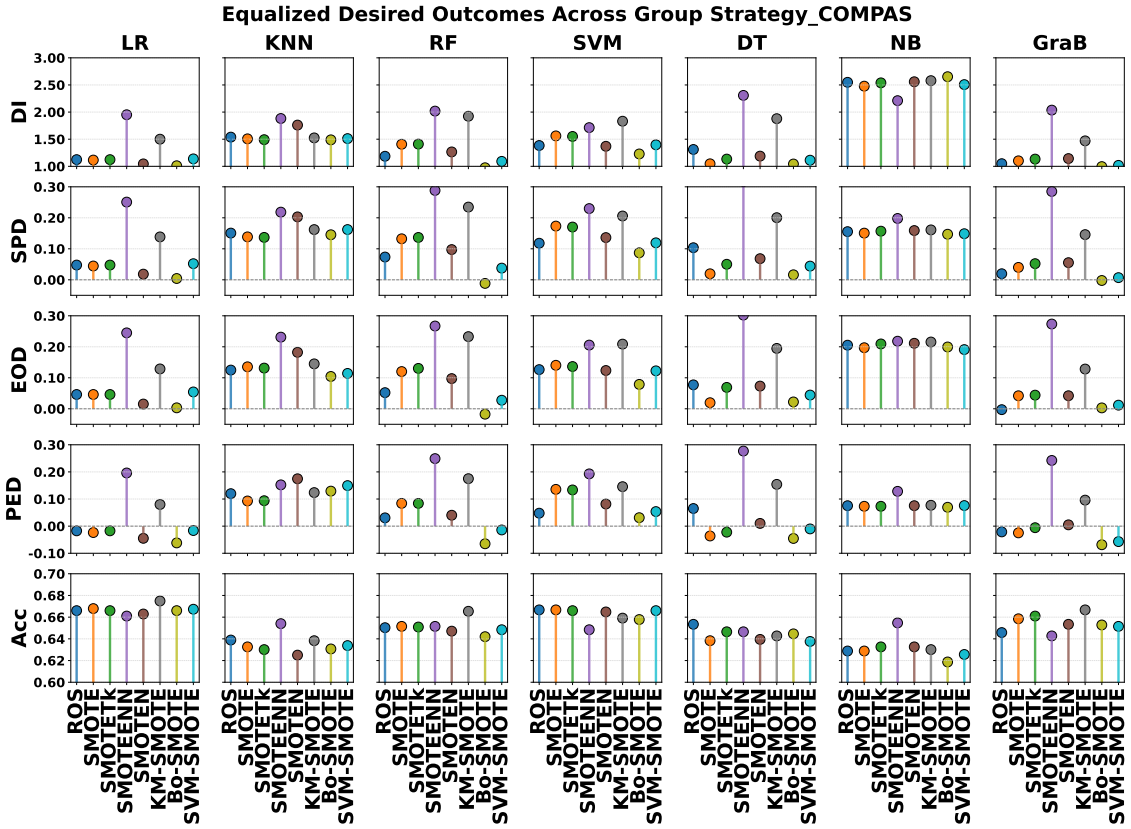

Figure 6.13: Equalized Desired Outcomes - COMPAS

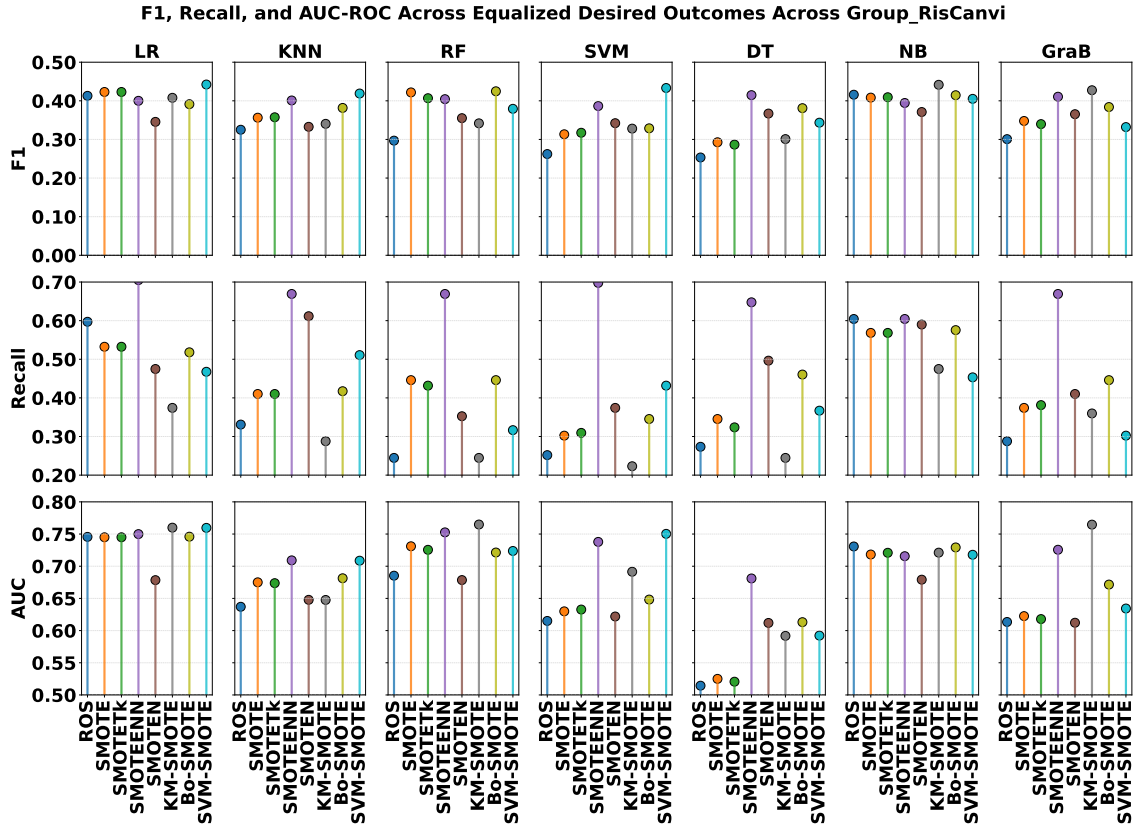

Figure 6.16: Equalized Desired Outcomes - RisCanvi

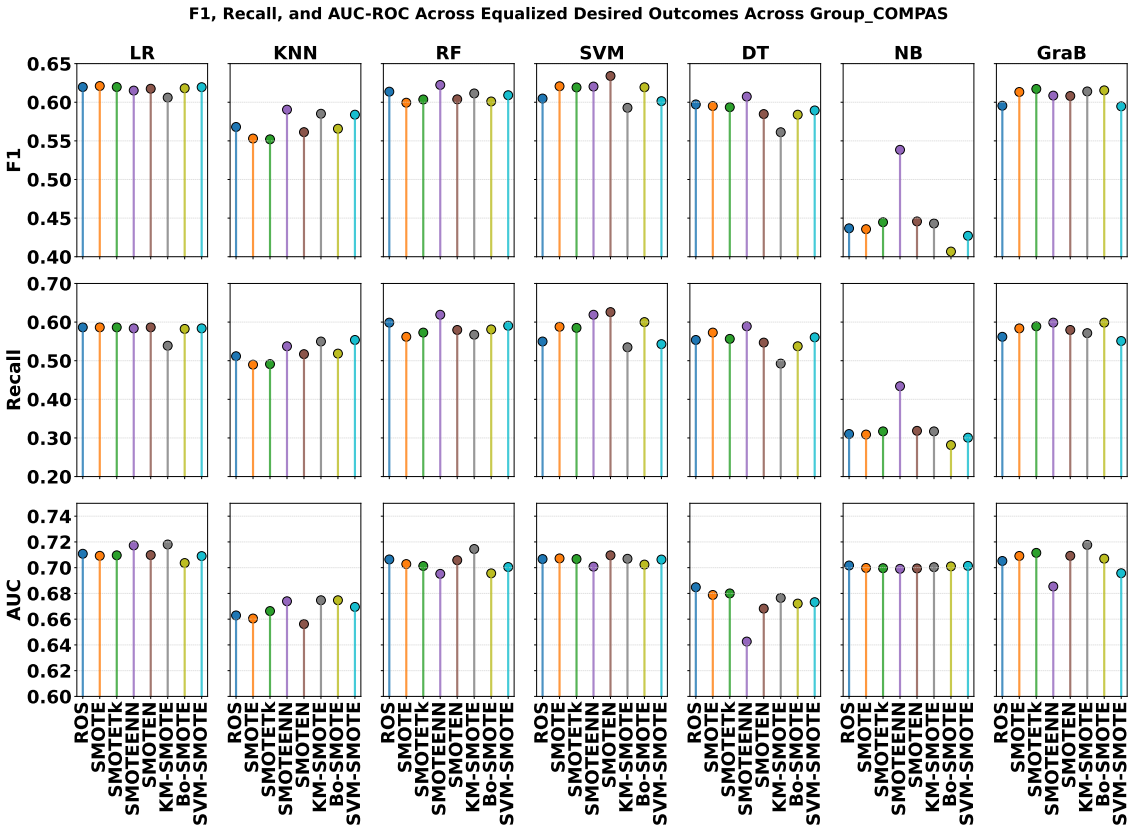

Figure 6.14: Equalized Desired Outcomes - COMPAS

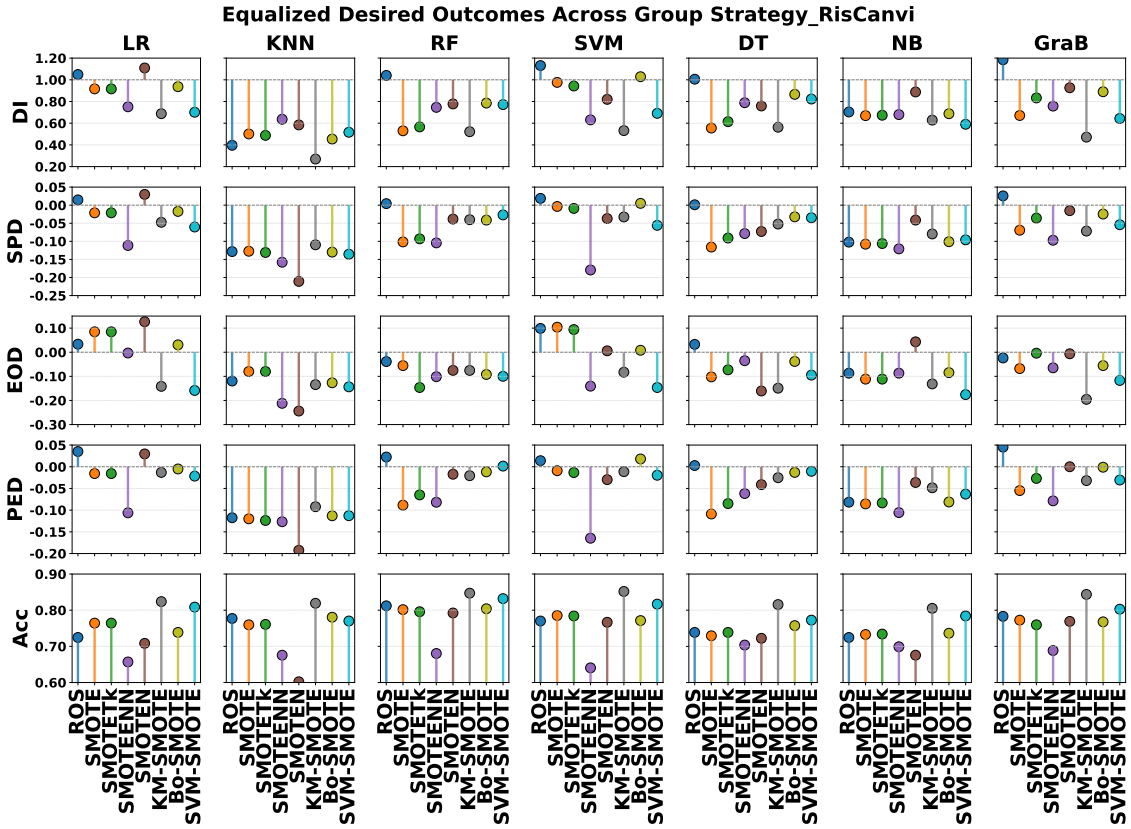

Figure 6.15: Equalized Desired Outcomes - RisCanvi

6.3.5 Summary of Comparative Findings

This comparative analysis reveals the differentiated effectiveness of oversampling strategies when applied across multiple fairness configurations and datasets. Traditional oversampling techniques provide a useful baseline, offering moderate gains in both fairness and classification accuracy; however, they often fail to resolve deeper systemic disparities embedded within the data. Oversampling based on sensitive attributes tends to yield stronger results in relatively balanced datasets like COMPAS but becomes less effective in datasets with pronounced group imbalances such as RisCanvi, where skewed distributions diminish its corrective power.

Strategies that equalize outcomes within the discriminated group show potential for targeted fairness improvements but risk excessive correction, particularly when using aggressive techniques like SMOTEENN, which can result in unstable or inflated fairness scores. In contrast, the approach that equalizes desired outcomes across both privileged and unprivileged groups proves to be the most reliable and broadly applicable. This method consistently delivers fairness metrics close to ideal targets while maintaining solid predictive performance across diverse classifiers.

Overall, the results emphasize that the success of an oversampling strategy is highly dependent on the dataset's structural characteristics, such as the extent of class imbalance, group distribution asymmetries, and classifier sensitivity. Therefore, careful alignment between resampling strategies and the underlying data properties is essential for producing models that are not only accurate but also equitable.

6.4 Discussion

This section revisits the study's research questions and provides an interpretive summary of the key findings, including the statistical relevance of observed trends.

6.4.1 How do fairness-aware oversampling techniques influence recidivism prediction accuracy and fairness across datasets with varying biases?

Fairness-aware oversampling methods exert a substantial, though context-sensitive, influence on both predictive accuracy and fairness metrics in recidivism forecasting tasks. In datasets such as COMPAS, which exhibit moderate demographic imbalance, these methods were generally effective in improving fairness while maintaining robust performance. For instance, combinations like *SVM-SMOTE with Logistic Regression (LR)* and *SMOTEN with SVM* achieved near-ideal fairness scores (e.g., DI $\approx$ 1.00, SPD $\approx$ 0.00), coupled with strong F1-scores exceeding 0.63 and AUC-ROC values approaching 0.72. These results indicate that, under certain conditions, fairness-aware oversampling can harmonize equity and effectiveness, especially when paired with classifiers that are less susceptible to data perturbations.

Conversely, in the RisCanvi dataset, characterized by more pronounced class imbalance and group skew, the outcomes were more varied. While configurations such as *SVM-SMOTE + LR* yielded reasonably balanced trade-offs (F1 $\approx$ 0.44, AUC-ROC $>$ 0.75), other pairings, including *SMOTEENN with Decision Tree (DT)*, demonstrated inflated fairness metrics (e.g., DI $>$ 2.5) and less reliable AUC-ROC scores. These disparities suggest that overly aggressive rebalancing may introduce synthetic noise, reducing generalization performance, particularly when sensitive group representation is limited.

Classifier choice further moderated these effects. Models like LR and GraB were consistently resilient across most fairness and accuracy measures, whereas Naive Bayes (NB) and DT were more susceptible to instability under data resampling. These variations reflect the need for careful coordination between oversampling strategy, dataset characteristics, and model complexity.

Insight: The impact of fairness-oriented oversampling is highly dependent on the underlying distribution of the data. In datasets with manageable bias levels,

such as COMPAS, these methods can significantly improve fairness without degrading accuracy. However, in more imbalanced datasets like RisCanvi, a cautious approach is needed to prevent overfitting or fairness distortion. Success hinges on selecting oversampling techniques that align with classifier sensitivity and the degree of structural bias in the dataset.

6.4.2 Which fairness-aware oversampling strategy is most effective in addressing systemic biases in recidivism datasets?

Among the evaluated strategies, the Equalized Desired Outcomes Across Groups approach consistently demonstrated the strongest capacity to mitigate systemic disparities in recidivism datasets. By balancing favorable outcomes across protected groups, this strategy directly aligns with fairness criteria such as Disparate Impact (DI), Statistical Parity Difference (SPD), Equal Opportunity Difference (EOD), and Predictive Equality Difference (PED). It achieved metrics closest to ideal fairness thresholds across both datasets, COMPAS and RisCanvi, while maintaining competitive predictive performance.

In the COMPAS dataset, this strategy, when implemented with classifiers like Logistic Regression (LR) and Gradient Boosting (GraB), resulted in substantial fairness gains without notable losses in accuracy. For instance, combinations such as *Bo-SMOTE + GraB* and *SMOTEN + SVM* reached nearly optimal fairness values (e.g., DI $\approx$ 1.00, SPD $\approx$ 0.00), while F1-scores remained above 0.63. Similarly, in the more challenging RisCanvi dataset, which exhibits stronger class and group imbalances, this strategy maintained its effectiveness. Configurations like *SVM-SMOTE + LR* and *SMOTEENN + RF* achieved F1-scores above 0.40 and AUC-ROC values above 0.75, reflecting its adaptability to more complex bias structures.

A key advantage of this method lies in its generalizability. It scaled effectively across various classifier architectures, both linear and non-linear, delivering consistent fairness outcomes. In contrast to targeted approaches like *Equalized Discrimi-*

nated Group Instances, which risk overcorrecting and inducing metric volatility, the Equalized Desired Outcomes strategy ensured population-level balance and mitigated overfitting.

Nonetheless, the approach is not without caveats. For example, classifiers such as Naive Bayes (NB), when paired with complex synthetic samplers, exhibited performance degradation, reflecting sensitivity to changes in data distribution. This highlights the need for thoughtful alignment between sampling method complexity and classifier architecture.

Conclusion: The Equalized Desired Outcomes Across Groups strategy proved to be the most robust and reliable oversampling method for mitigating systemic bias, achieving consistent fairness outcomes and maintaining predictive validity across diverse datasets and classifiers.

6.4.3 How do different classifiers perform when paired with fairness-aware oversampling methods, and what trade-offs arise between fairness and predictive performance?

Classifier performance varied significantly when combined with fairness-aware oversampling methods, often revealing trade-offs between equity and accuracy. Logistic Regression (LR), known for its simplicity and interpretability, consistently performed well in terms of fairness, particularly when paired with oversampling techniques like Bo-SMOTE, SMOTEN, or SVM-SMOTE. Across both COMPAS and RisCanvi datasets, LR-based models frequently achieved parity-level fairness (DI $\approx$ 1.00, SPD $\approx$ 0.00), alongside solid predictive performance (F1-scores and AUC-ROC values remained competitive). This suggests that LR's relatively low model complexity enables it to accommodate the controlled variance introduced by synthetic sampling.

Random Forest (RF) and other ensemble models, such as Gradient Boosting (GraB), often delivered stronger raw classification metrics (e.g., Recall $>$ 0.66 in

certain SMOTEENN scenarios), but their fairness scores were more volatile. These models have a tendency to overfit on oversampled training data, reinforcing pre-existing biases when exposed to aggressive synthetic sampling.

Support Vector Machines (SVM) showed mixed outcomes. While the model performed well with sampling strategies like SMOTEN and SVM-SMOTE, its performance degraded notably with clustering-based methods such as KMeans-SMOTE, likely due to increased noise sensitivity. Similarly, Naive Bayes (NB) often experienced lower F1-scores, particularly when faced with oversampled distributions in high-dimensional feature spaces, even though its fairness scores were occasionally moderate.

Trade-off Insight: The interplay between classifier architecture and oversampling strategy shapes both fairness and performance outcomes. Linear models like LR offer balanced results and are less prone to fairness distortion, whereas ensemble models often enhance accuracy at the cost of equitable outcomes. Moreover, aggressive sampling techniques may inflate fairness metrics while negatively impacting generalization (as indicated by lower AUC-ROC or precision). These results underscore the importance of aligning classifier complexity with data augmentation strategies to strike an effective balance between fairness and predictive strength.

6.4.4 Can fairness-aware oversampling techniques generalize effectively across recidivism datasets with different levels of bias and class imbalance?

Fairness-oriented oversampling methods exhibit a degree of cross-dataset generalizability, though their effectiveness is often constrained by the specific characteristics of each dataset. In the case of the COMPAS dataset, characterized by moderate class imbalance and demographic disparities, approaches such as *Equalized Desired Outcomes Across Groups* combined with classifiers like Logistic Regression (LR) or Gradient Boosting (GraB) consistently achieved parity-level fairness metrics (DI $\approx$ 1.00, SPD $\approx$ 0.00), while maintaining strong predictive results (F1 $>$ 0.63,

AUC-ROC > 0.71).

By contrast, the RisCanvi dataset posed greater challenges due to its pronounced class imbalance and heavily skewed subgroup distributions, particularly in terms of positive outcomes among disadvantaged groups. Under these conditions, some techniques that performed well in COMPAS, such as SMOTEENN, led to fairness inflation (e.g., DI > 2.5) or degradation in predictive quality (e.g., reduced AUC-ROC), indicating potential overfitting or overcompensation. Despite these limitations, specific combinations, such as *SVM-SMOTE + LR* and *SMOTEN + GraB*, were still able to deliver balanced fairness and performance metrics under more adverse conditions.

Generalization Insight: Oversampling strategies like Bo-SMOTE and RandomOverSampler, when paired with robust classifiers such as LR or GraB, show the most promise for generalization across datasets. However, their effectiveness diminishes significantly in the presence of severe imbalance or highly skewed group distributions. These findings highlight the importance of tailoring oversampling strategies to the data's specific structure and the classifier's sensitivity to distributional shifts. Without such contextual alignment, fairness-aware interventions risk either ineffectiveness or instability, particularly problematic in domains like criminal justice, where model outputs carry real-world consequences.

Guidance for Generalizing to New Datasets: The applicability of fairness-aware oversampling techniques is closely tied to the underlying statistical properties of the dataset. Techniques like *Bo-SMOTE* and *Equalized Desired Outcomes Across Groups* tend to perform well in datasets that exhibit both significant class imbalance and group disparity, particularly when the protected group has a low rate of favorable outcomes. Conversely, simpler methods like *RandomOverSampler* remain viable in contexts with moderate demographic skew and more balanced outcome distributions.

The dimensionality and richness of the dataset also play a crucial role. In feature-rich datasets such as COMPAS, complex oversampling techniques are more likely to

generate useful synthetic data without substantially harming generalization. However, in sparse or low-dimensional datasets like RisCanvi, aggressive resampling can introduce harmful noise. As such, practitioners should assess class distribution, subgroup representation, and feature complexity before applying oversampling strategies. In high-imbalance scenarios, fairness-driven approaches should be considered, but with careful attention to the dataset's capacity to support robust synthetic data generation.

6.4.5 Statistical Significance and Confidence Intervals

To ensure the reliability of our results, we computed 95% confidence intervals (CIs) for all fairness metrics and accuracy scores across oversampling strategies and models. Due to space constraints, the full statistical breakdown is not included here but is accessible in the supplementary notebooks: "Confidence Interval for RisCanvi", "Confidence Interval for COMPAS", and "RisCanvi and COMPAS Statistical Analysis". Table 6.4 presents mean scores and confidence intervals for the Equalized Discriminated Group Instances oversampling strategy on both datasets.

Logistic Regression (LR) demonstrated the most stable and consistent fairness outcomes across both datasets. In COMPAS, it achieved a DI of 1.25 with a 95% CI of [0.98, 1.52], SPD of 0.075 [0.0076, 0.1432], and EOD of 0.073 [0.0073, 0.1389], all close to ideal, with CIs that encompass the parity thresholds, indicating no significant unfairness. Similar results were observed in RisCanvi, where LR's DI was 0.88 [0.75, 1.01] and other fairness metrics similarly hovered near zero, reinforcing its reliability. Additionally, LR delivered competitive accuracy in both datasets (0.6665 for COMPAS, 0.7487 for RisCanvi).

Gradient Boosting (GraB) also performed well, albeit with slightly broader intervals. On COMPAS, GraB achieved a DI of 1.24 [0.95, 1.54] and SPD of 0.075 [–0.0052, 0.156], indicating proximity to fairness though with less certainty. In RisCanvi, GraB yielded a DI of 0.80 [0.62, 0.98] and SPD of –0.042 [–0.075, –0.010], suggesting modest bias but overall acceptable bounds. Its accuracy remained strong

at 0.6541 (COMPAS) and 0.7733 (RisCanvi).

In contrast, K-Nearest Neighbors (KNN) and Naïve Bayes (NB) revealed significant fairness challenges. KNN's DI values varied greatly, 0.48 [0.39, 0.58] in RisCanvi and 1.59 [1.46, 1.71] in COMPAS, indicating inconsistency and potential unreliability. NB was particularly problematic in COMPAS, with a DI of 2.51 [2.40, 2.62], far from the ideal. While NB achieved high accuracy, its fairness deviations suggest it may amplify pre-existing biases.

Support Vector Machine (SVM) and Random Forest (RF) offered strong classification metrics but mixed fairness results. In COMPAS, SVM achieved DI = 1.50 [1.34, 1.67], SPD = 0.155 [0.115, 0.195], and EOD = 0.143 [0.106, 0.179], all statistically significant and outside the fairness thresholds. However, these results did not fully translate to RisCanvi, where fairness metrics often excluded the ideal values, indicating disparities. Nonetheless, both classifiers posted high accuracies in RisCanvi (SVM: 0.7733, RF: 0.7957), making them suitable in contexts prioritizing predictive strength over fairness.

Decision Tree (DT) yielded intermediate outcomes. In COMPAS, it had a DI of 1.38 [0.99, 1.77] and SPD of 0.102 [0.015, 0.190], both close to parity. On RisCanvi, DT showed slightly improved fairness (DI = 0.75 [0.61, 0.88]) but at the cost of lower accuracy. While not the best in any single category, DT delivered reasonable trade-offs.

Summary Insight: These results illustrate that not all high-accuracy models are fair, and not all fair models offer top-tier predictive performance. While classifiers like RF and NB excel in accuracy, they may suffer from substantial bias. In contrast, simpler models like LR manage to balance fairness and accuracy effectively, making them preferable for fairness-critical applications.

Table 6.4: Mean values and 95% confidence intervals (CI) for fairness metrics and accuracy across multiple classifiers under the Equalized Discrimination Group Instances oversampling strategy, evaluated on the COMPAS and RisCanvi recidivism prediction datasets. Metrics include Disparate Impact (DI), Statistical Parity Difference (SPD), Equal Opportunity Difference (EOD), Predictive Equality Difference (PED), and Accuracy (Acc).

Metrics	Model	RisCanvi			COMPAS		
		Mean	Lower CI	Upper CI	Mean	Lower CI	Upper CI
DI	LR	0.8833	0.7521	1.0145	1.2508	0.9835	1.5181
	KNN	0.4806	0.3863	0.5750	1.5877	1.4636	1.7118
	RF	0.7172	0.5714	0.8630	1.4092	1.0938	1.7247
	SVM	0.8440	0.6678	1.0201	1.5043	1.3380	1.6705
	DT	0.7468	0.6138	0.8797	1.3775	0.9896	1.7654
	NB	0.6900	0.6166	0.7633	2.5091	2.3995	2.6187
	GraB	0.7972	0.6173	0.9770	1.2431	0.9480	1.5382
SPD	LR	-0.0293	-0.0664	0.0077	0.0754	0.0076	0.1432
	KNN	-0.1413	-0.1672	-0.1153	0.1646	0.1393	0.1898
	RF	-0.0552	-0.0884	-0.0220	0.1236	0.0409	0.2063
	SVM	-0.0367	-0.0890	0.0156	0.1550	0.1146	0.1954
	DT	-0.0596	-0.0908	-0.0283	0.1024	0.0145	0.1903
	NB	-0.0944	-0.1148	-0.0740	0.1596	0.1459	0.1732
	GraB	-0.0668	-0.0750	-0.0105	0.0753	-0.0052	0.1558
EOD	LR	0.0071	-0.0807	0.0949	0.0731	0.0073	0.1389
	KNN	-0.1425	-0.1913	-0.0938	0.1460	0.1113	0.1808
	RF	-0.0856	-0.1130	-0.0582	0.1137	0.0322	0.1953
	SVM	-0.0074	-0.0952	0.0804	0.1427	0.1060	0.1794
	DT	-0.0776	-0.1309	-0.0243	0.1003	0.0181	0.1825
	NB	-0.0932	-0.1458	-0.0406	0.2058	0.1979	0.2137
	GraB	-0.0668	-0.1203	-0.0133	0.0677	-0.0098	0.1452
PED	LR	-0.0141	-0.0501	0.0219	0.0119	-0.0595	0.0832
	KNN	-0.1248	-0.1493	-0.1003	0.1294	0.1056	0.1533
	RF	-0.0326	-0.0666	0.0014	0.0730	-0.01130	0.1572
	SVM	-0.0270	-0.0754	0.0214	0.1027	0.0548	0.1506
	DT	-0.0428	-0.0757	-0.0100	0.0490	-0.0455	0.1435
	NB	-0.0732	-0.0919	-0.0544	0.0810	0.0649	0.0971
	GraB	-0.0224	-0.0538	0.0091	0.0207	-0.0650	0.1064
Acc	LR	0.7487	0.7035	0.7939	0.6665	0.6631	0.6699
	KNN	0.7431	0.6848	0.8015	0.6354	0.6281	0.6427
	RF	0.7957	0.7537	0.8376	0.6509	0.6453	0.6565
	SVM	0.7733	0.7224	0.8243	0.6619	0.6565	0.6674
	DT	0.7472	0.7182	0.7762	0.6436	0.6392	0.6481
	NB	0.7364	0.7014	0.7714	0.6314	0.6228	0.6402
	GraB	0.7733	0.7368	0.8098	0.6541	0.6475	0.6606

6.5 Recommended Fairness-Aware Pipeline for Recidivism Prediction

Based on our comparative analysis across both the COMPAS and RisCanvi datasets, we propose the most effective combinations of oversampling techniques, classification models, and fairness-oriented strategies for recidivism prediction. These selections are grounded in their consistent ability to achieve a balanced trade-off between fairness and predictive accuracy under varying data characteristics.

Table 6.5 outlines the leading configurations identified throughout the experiments. These are organized by oversampling method, classifier, and fairness strategy. The primary options represent those that produced the most reliable results across both datasets, while secondary options are offered as viable substitutes in scenarios where the primary selections may not be applicable or optimal for specific contexts.

Table 6.5: Top Recommended Components for Fairness-Aware Recidivism Prediction

Category	**Primary Choice**	**Secondary Choice**
Oversampling Technique	Borderline-SMOTE	SMOTEN
Classifier	Logistic Regression (LR)	Gradient Boosting (GraB)
Fairness Strategy	Equalized Desired Outcomes	Equalized Discriminated Groups

These findings underscore that no single configuration guarantees optimal results in all settings. However, combinations like *SMOTEN + LR* or *Bo-SMOTE + GraB*, when applied under the Equalized Desired Outcomes strategy, offer a robust starting point. Practitioners should adapt these recommendations to the specific properties of their datasets, particularly with respect to class imbalance, group skew, and the complexity of the feature space. Moreover, post-processing adjustments such as threshold calibration may further enhance fairness-performance trade-offs.

6.5.1 Recommended Pipeline Configuration

The optimal pipeline combines Borderline-SMOTE with Logistic Regression, aligned with the Equalized Desired Outcomes Across Groups approach. This setup consistently achieved fairness metrics close to ideal values across both datasets. Disparate Impact remained in the narrow range of 0.99 to 1.01, and Statistical Parity Difference was tightly constrained between –0.003 and 0.01. On the predictive side, it maintained strong performance, with F1-scores spanning from 0.44 to 0.63 and AUC-ROC values between 0.71 and 0.76. Additionally, 95% confidence intervals for both fairness and accuracy metrics were compact, reflecting the model's robustness and consistency. These qualities make this configuration particularly suitable for deployment in high-stakes, fairness-sensitive settings like criminal justice risk assessments, where both equitable treatment and dependable predictions are paramount.

6.5.2 Recommendations in Practice

The experimental findings reveal that oversampling techniques interact differently with classifiers and dataset characteristics, producing both benefits and risks. For example, applying SVM-SMOTE to the highly imbalanced RisCanvi dataset caused predictive accuracy to collapse (F1-scores ≈ 0.01), while combining SMOTEENN with Decision Trees yielded Disparate Impact values exceeding 2.7, indicating overcorrection. Naive Bayes models also performed poorly across datasets, rarely exceeding an F1-score of 0.43, likely due to their sensitivity to distributional shifts introduced by synthetic resampling. These observations highlight the need for practitioners to view resampling not as a universal solution but as a context-dependent tool whose effectiveness is shaped by both model choice and dataset properties.

From these limitations, several practical recommendations follow. First, oversampling methods should always be aligned with dataset imbalance conditions: in severely skewed datasets such as RisCanvi, simple approaches like RandomOverSampler or SMOTEN, paired with stable classifiers, tend to yield more dependable results. In contrast, more complex oversamplers can be useful in moderately im-

balanced datasets, provided they are combined with models robust to synthetic variability. Second, oversampling should be integrated into a broader fairness-aware pipeline that includes sensitivity analysis, and context-specific validation rather than being applied in isolation. Finally, in high-stakes domains such as criminal justice, fairness interventions must be context-aware, legally defensible, and subject to continuous monitoring after deployment.

To support implementation, Table 6.6 outlines recommended oversampler–classifier pairings tailored to different dataset scenarios. For datasets with moderate class and subgroup imbalance (e.g., COMPAS), strategies such as Equalized Desired Outcomes Across Groups paired with Borderline-SMOTE, SMOTEN, or SVM-SMOTE perform well when used with robust classifiers like Logistic Regression or Gradient Boosting. In contrast, highly imbalanced datasets with limited feature diversity (e.g., RisCanvi) are better served by simpler oversamplers such as RandomOverSampler, especially when combined with models resilient to sampling-induced variability. Where fairness takes precedence over accuracy, Equalized Desired Outcomes remains the preferred choice; however, when predictive performance is prioritised, methods such as SMOTEENN combined with ensemble classifiers may be suitable, provided fairness metrics are regularly audited.

In summary, these recommendations provide a framework for making informed methodological choices that balance fairness and predictive performance, enabling more responsible deployment of fairness-aware recidivism prediction models.

6.5.3 Comparison with Fairness-Aware Learning Techniques

To assess the relative performance of our oversampling-based fairness strategies, we conducted a comparison with a range of established fairness-improving methods across different stages of the machine learning pipeline. This includes preprocessing methods from the AIF360 framework, such as the Disparate Impact Remover (DIR) and Reweighing (Re); in-processing techniques like Exponentiated Gradient Reduction (EGR) and Adversarial Learning (AL); and post-processing methods such as

Table 6.6: Recommended Oversampler–Classifier Combinations for Different Dataset Conditions

Dataset Condition	**Recommended Oversampler(s)**	**Recommended Classifier(s)**
Moderate class and group imbalance (e.g., COMPAS)	Equalized Desired Outcomes Across Groups + Borderline-SMOTE, SMOTEN, or SVM-SMOTE	Logistic Regression, Gradient Boosting
Severe group and class imbalance (e.g., RisCanvi)	Equalized Desired Outcomes Across Groups or RandomOverSampler (with cautious tuning)	Logistic Regression, Gradient Boosting; avoid Naive Bayes
Limited features or sparse datasets	RandomOverSampler or SMOTEN (avoid complex oversamplers like KMeans-SMOTE)	Logistic Regression, Support Vector Machine
High-dimensional feature space	Borderline-SMOTE or SMOTEENN (with fairness-aware strategy)	Gradient Boosting, Random Forest (monitor fairness trade-off)
Fairness is top priority	Equalized Desired Outcomes Across Groups + SVM-SMOTE or SMOTEN	Logistic Regression
Accuracy is top priority	RandomOverSampler or SMOTEENN (without fairness constraint)	Random Forest, Gradient Boosting

Equalized Odds Optimization (EO) and Reject Option Classification (ROC). These techniques offer comprehensive coverage of fairness interventions at the data, model, and decision levels, providing a rigorous benchmark for evaluating our proposed approach.

Table 6.7 summarizes the comparative results on both the COMPAS and RisCanvi datasets, measured across four key fairness indicators, *Statistical Parity Difference (SPD)*, *Disparate Impact (DI)*, *Equal Opportunity Difference (EOD)*, and *Predictive Equality Difference (PED)*, as well as overall classification accuracy (Acc).

As shown in Table 6.7, models trained without any fairness intervention exhibit substantial bias, particularly in the COMPAS dataset, where the Statistical Parity Difference (SPD) reaches 0.25 and the Disparate Impact (DI) is 2.13. Preprocessing methods such as Reweighing and Disparate Impact Remover (DIR) offer marginal improvements. For instance, DIR reduces DI to 0.7764 on RisCanvi but increases

Table 6.7: Comparison of fairness and performance metrics across diverse fairness-enhancing strategies on RisCanvi and COMPAS datasets

Techniques	RisCanvi					COMPAS				
	SPD	DI	EOD	PED	Acc	SPD	DI	EOD	PED	Acc
None	0.059	0.8170	**0.011**	0.045	0.72	0.25	2.13	0.025	0.18	**0.68**
Re	0.046	0.8658	**0.011**	0.039	0.71	0.097	1.243	0.094	0.031	0.66
DIR	0.079	0.7764	0.056	0.067	0.71	0.284	2.092	0.281	0.211	0.67
EGR	0.082	0.785	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**	**0.9542**	0.033	**0.004**	0.70	0.040	1.125	0.014	**0.015**	0.64
ROC	0.079	0.777	0.056	0.066	0.71	0.270	2.100	0.287	0.183	**0.68**
Bo-SMOTE+LR (Ours)	-0.017	0.9366	0.031	-0.005	**0.74**	**0.004**	**1.0104**	**0.003**	-0.062	0.67

Equal Opportunity Difference (EOD) and Predictive Equality Difference (PED) to 0.281 and 0.211 on COMPAS, respectively. Reweighing provides only modest fairness enhancements and lowers the accuracy on COMPAS to 0.66.

In-processing techniques yield mixed results. Exponentiated Gradient Reduction (EGR) achieves DI values of 0.785 on RisCanvi and 1.093 on COMPAS, but these come with increased EOD and PED, for example, EOD = 0.135 and PED = 0.083 on RisCanvi. Adversarial Learning (AL) produces a relatively low SPD (0.064) but results in a high DI of 1.237 on RisCanvi and reduced accuracy on COMPAS.

Post-processing strategies demonstrate more focused improvements. Equalized Odds Optimization (EO) achieves the best PED on RisCanvi (0.004) and competitive EOD on COMPAS (0.014), but it also reduces overall accuracy to 0.64. Reject Option Classification (ROC) shows moderate DI and SPD on RisCanvi but retains significant bias in COMPAS, where EOD rises to 0.287 and DI to 2.100.

By comparison, the Equalized Desired Outcomes approach, implemented through Borderline-SMOTE combined with Logistic Regression, achieves the most consistent and balanced results across datasets. On COMPAS, it yields an SPD of 0.004, a DI close to parity at 1.0104, and a PED of –0.062, all while maintaining accuracy at 0.67. In RisCanvi, it provides the best accuracy (0.74), the lowest SPD (–0.017), and a near-ideal DI of 0.9366.

These outcomes indicate that while in- and post-processing techniques offer theoretically grounded fairness interventions, their real-world effectiveness may vary depending on dataset properties. In contrast, the Equalized Desired Outcomes

strategy demonstrates superior consistency across both fairness and accuracy metrics. Its transparent design and compatibility with common classifiers make it a practical choice for applications where fairness is a primary concern.

6.6 Ethical and Societal Implications

6.6.1 Addressing Structural Inequities in Criminal Justice

Our findings affirm that fairness-aware oversampling, especially the "Equalized Desired Outcomes Across Groups" strategy, when implemented with techniques like Borderline-SMOTE, can significantly reduce disparity across fairness metrics. In the criminal justice domain, where tools such as COMPAS have historically displayed racially biased outcomes, these methods offer a viable approach for reducing algorithmic harm. By rebalancing class and demographic distributions, these techniques help counteract the perpetuation of historical inequalities embedded in risk prediction systems.

6.6.2 Balancing Fairness and Predictive Accuracy

Although fairness-enhancing approaches can improve equity, they also introduce potential downsides. For example, in the RisCanvi dataset, certain configurations, such as SMOTEENN with Logistic Regression, resulted in Disparate Impact values exceeding 3.0 (see Section 6.3.3, Figure 6.11), suggesting excessive bias correction. Such outcomes risk impairing model generalizability, potentially increasing the rate of incorrect classifications. In sensitive contexts like parole or sentencing, this could lead to ethically questionable decisions. These findings highlight the need for a careful balance between fairness objectives and predictive reliability during model deployment.

6.6.3 Model Transparency and Responsibility

Classifiers such as Logistic Regression (LR) and Gradient Boosting (GraB) demonstrated strong performance while remaining interpretable (Section 6.4.3). Their transparency makes them preferable in applications requiring accountability, such as judicial decision-making, where it is essential that outputs can be understood, audited, and challenged if necessary. In contrast, opaque or "black-box" models may obscure bias sources and hinder trust, making them ethically less desirable for high-stakes scenarios.

6.6.4 Contextual Fairness Across Jurisdictions

The performance of fairness interventions varied significantly between datasets. While oversampling techniques improved fairness on the U.S.-based COMPAS dataset, results were more inconsistent on the highly imbalanced Catalan dataset, RisCanvi (see Section 6.4.4). This divergence suggests that effective deployment of fairness-aware methods requires sensitivity to regional, legal, and cultural contexts. What works well in one setting may be ineffective or counterproductive in another without tailored adaptation and validation.

6.6.5 Beyond Technical Solutions

While fairness-aware oversampling contributes to improving algorithmic equity, its societal impact is limited without complementary structural reforms. Algorithms represent only one component within broader justice systems. Over-reliance on technical fixes may mask deeper societal issues such as income inequality, racial profiling, or systemic bias in law enforcement (Section 6.6). For AI-based tools to genuinely promote ethical justice, they must be deployed alongside broader policy interventions that address the root causes of inequality, not just their computational manifestations.

6.7 Limitations

6.7.1 Constraints of Dataset Scope

This work utilizes two datasets: COMPAS and RisCanvi. While COMPAS is a widely used benchmark in algorithmic fairness research, it is rooted in the U.S. legal system and may not generalize to other socio-legal contexts. RisCanvi offers a valuable non-U.S. perspective from Catalonia but presents significant challenges due to pronounced class and group imbalances (e.g., 464 positive vs. 2,391 negative outcomes; see Table 6.1). These disparities limit the applicability of findings to datasets with different structural properties or institutional settings.

6.7.2 Classifier-Dependent Effectiveness of Oversampling

The impact of oversampling strategies was not uniform across classifiers. Models such as Naive Bayes consistently struggled with synthetic data generated by more aggressive techniques like SMOTEENN, often resulting in reduced performance. In contrast, classifiers like Logistic Regression (LR) and Gradient Boosting (GraB) demonstrated more robust and reliable trade-offs between fairness and predictive accuracy (refer to Section 6.4.3). These variations suggest that oversampling efficacy is influenced by the classifier's capacity to handle distributional shifts, emphasizing the importance of method–model compatibility.

6.7.3 Risk of Overcompensation and Metric Ambiguities

Some oversampling configurations led to excessively high fairness metrics (e.g., $DI > 2.0$), indicating potential overcorrection effects (see Section 6.3.3, Figure 6.11). Moreover, group-level fairness metrics, such as DI, SPD, EOD, and PED, while useful for summarizing disparities, may fail to capture individual-level fairness violations or unintended consequences of model deployment. For instance, models that appear fair on average might still produce harmful decisions for specific subgroups. Future investigations should incorporate individual fairness frameworks and assess

real-world impact.

6.7.4 Challenges in Data Representation and Synthetic Sampling

Techniques like SMOTE assume that feature spaces are sufficiently rich to generate meaningful synthetic instances. However, in high-dimensional or low-sample contexts, such as the RisCanvi dataset, these assumptions may break down, resulting in noisy synthetic data and degraded model performance (see Section 6.5.2). Additionally, hybrid resampling methods such as SMOTEENN yielded unstable outcomes across different classifiers, particularly in terms of recall and AUC-ROC (Section 6.3.2, Figures 6.6 and 6.8), revealing limitations in their generalizability across diverse data structures.

6.8 Summary

This chapter comprehensively evaluated the impact of fairness-aware oversampling strategies on both predictive performance and fairness outcomes in recidivism prediction. Using two datasets, COMPAS and RisCanvi, with differing levels of bias and imbalance, the study systematically compared four major oversampling strategies: Traditional Oversampling, Sensitive Attribute-Based Sampling, Equalized Discriminated Group Instances, and Equalized Desired Outcomes Across Groups.

The findings indicate that while traditional methods offer baseline improvements, more advanced strategies, particularly Equalized Desired Outcomes, yield the most consistent and equitable results across both datasets. Logistic Regression and Gradient Boosting emerged as the most robust classifiers, striking a reliable balance between fairness metrics (e.g., DI ≈ 1.00, SPD ≈ 0.00) and predictive performance (F1 > 0.63, AUC-ROC > 0.71). However, classifier-sampler compatibility and dataset characteristics (e.g., class imbalance, feature richness) heavily influenced outcomes. Oversampling methods such as SMOTEENN, though effective in some contexts,

posed risks of fairness overcorrection or performance instability, especially in imbalanced datasets like RisCanvi.

Moreover, comparative analysis with AIF360 fairness-enhancing methods demonstrated that the proposed oversampling-based pipelines can achieve parity with or outperform established techniques, with greater simplicity and transparency. Ethical and societal implications were also discussed, emphasizing the importance of aligning technical interventions with interpretability, legal accountability, and domain-specific fairness requirements.

Overall, this chapter underscores that fairness-aware oversampling is a powerful but context-sensitive tool. Its effectiveness hinges on thoughtful alignment with classifier choice, dataset properties, and the fairness-performance trade-offs specific to real-world applications like criminal justice.

Publication(s) Arising from this Chapter

The work presented in this chapter has been published (or submitted) in the following outlets:

1. Michael Mayowa Farayola, Malika Bendechache, Saber Takfarinas, et al. (2025). "Investigating Fairness-Aware Oversampling Strategies and Techniques Across Diverse Machine Learning Algorithms for Recidivism Prediction". In: *Minds and Machines* 35.3, p. 37

These publications reflect the main contributions of this chapter and provide further technical details, extended results, and peer-reviewed validation of the methods and findings.

Chapter 7

Intersectional Fairness: Auditing Compound Bias in Algorithmic Decisions

7.1 Introduction

Artificial intelligence (AI) is increasingly embedded in high-stakes decision-making across domains such as criminal justice, finance, and health insurance (Montani and Striani 2019; Khreisat et al. 2024; Fabris et al. 2025). These systems are praised for their ability to enhance efficiency, consistency, and scalability in decisions that directly impact human lives (R. Berk and Bleich 2014; Caroline Wang et al. 2022; Pelegrina, Couceiro, and Duarte 2023). However, alongside these benefits, serious concerns have emerged regarding their potential to perpetuate or exacerbate structural inequities, especially when multiple protected attributes intersect (Kearns et al. 2018).

A key conceptual foundation for these concerns is *intersectionality*, first articulated in Black feminist theory (Crenshaw 2022; Al-Faham, Davis, and Ernst 2019; Ovalle et al. 2023). Intersectionality highlights how individuals situated at the intersection of multiple marginalized identities, such as Black women or elderly Latina

patients, may experience unique and compounded forms of disadvantage that remain invisible when attributes like race or gender or age are considered in isolation. In algorithmic contexts, this means a model may appear fair with respect to race or gender independently, yet still produce disproportionate harm to those at their intersection. Kearns Michael (Kearns et al. 2018) characterize this phenomenon as *fairness gerrymandering*, where models satisfy fairness constraints for large groups while concealing harms to smaller, more vulnerable subgroups.

Hence, this chapter addresses these limitations by introducing a framework for auditing AI systems through an intersectional lens. By extending fairness assessments beyond group-level averages and incorporating disaggregated subgroup analysis, we aim to uncover disparities that aggregate metrics often obscure. Our empirical evaluation is centered on two benchmark datasets, COMPAS and Adult Income, where intersectional subgroups are defined primarily along race–gender dimensions. In addition, a health insurance dataset is employed as a *validation exercise* and real-world application, used to demonstrate the generalizability of fairness-enhancing interventions across domains by incorporating intersections of race, gender, and age.

The chapter therefore pursues two core objectives: (1) to surface the limitations of conventional fairness metrics when applied to real-world AI systems, and (2) to empirically assess the performance of fairness-enhancing interventions when evaluated through an intersectional framework.

7.2 Motivation and Problem Statement

Despite progress in fairness auditing, the majority of algorithmic evaluations still examine sensitive attributes, such as race, gender, or age, in isolation (Roy, Horstmann, and Ntoutsi 2023; Ovalle et al. 2023). This one-dimensional perspective obscures the reality that individuals can simultaneously occupy multiple marginalized identities. As a result, AI systems may pass standard fairness evaluations on single attributes while still producing compounded harms for intersectional subgroups.

For instance, a risk assessment tool may demonstrate similar false positive rates

across racial groups and across gender groups independently, yet consistently misclassify Black women at disproportionately higher rates. Similarly, in health insurance, algorithmic triage may appear unbiased along singular dimensions such as gender or age, yet systematically fails older Latina patients due to compounded under representation and systemic bias embedded in training data and care delivery.

These intersectional harms are not mere statistical anomalies; they reflect deeper structural inequities encoded within data sources and decision-making pipelines. Failing to detect and address these harms undermines the ethical integrity of AI systems, especially in high-stakes domains where decisions influence freedom, health, or access to essential resources.

This chapter therefore advocates for a paradigm shift in fairness evaluation: from isolated, single-attribute assessments to intersectional analyses that more accurately reflect the lived experiences of those most vulnerable to algorithmic bias. By doing so, we aim to move beyond superficial fairness guarantees and toward a more contextually grounded, justice-aware approach.

To address this challenge, this study contributes to the algorithmic fairness literature by evaluating the performance of integrated fairness-enhancing techniques when assessed through an intersectional lens. Prior research has primarily focused on single protected attributes and aggregate metrics, which risks obscuring disparities that affect individuals with intersecting marginalized identities (Crenshaw 2022; Mehrabi et al. 2021; Hanna et al. 2020a).

We investigate whether mitigation strategies that explicitly account for intersectional groupings can improve fairness outcomes without significantly compromising predictive performance. We conduct a comparative analysis of individual and combined fairness interventions across two benchmark datasets, Adult (Census Income) and COMPAS (recidivism), evaluating results using four fairness metrics: Statistical Parity Difference (SPD), Disparate Impact (DI), Equal Opportunity Difference (EOD), and Predictive Equality Difference (PED), alongside predictive accuracy.

Central Hypothesis: *Integrated fairness-enhancing techniques that incorpo-*

rate intersectional protected attributes (e.g., race and gender jointly) can achieve improved fairness outcomes while maintaining competitive predictive accuracy.

To empirically assess this hypothesis, we pose the following research questions:

- **RQ1**: Do integrated fairness-enhancing models reduce disparities across intersectional subgroups (e.g., race × gender), and do aggregate metrics obscure these disparities?
- **RQ2**: What trade-offs arise between fairness metrics (e.g., SPD, DI, EOD, PED) and predictive accuracy when incorporating intersectionality?
- **RQ3**: Which fairness mitigation strategies (individual vs. integrated) are most effective for improving subgroup fairness while maintaining accuracy?
- **RQ4**: How do fairness mitigation strategies perform differently across domains (e.g., COMPAS, Adult Income, Irish Health Insurance)?
- **RQ5**: How does the stage of intervention (pre-processing, in-processing, or post-processing) influence fairness outcomes?
- **RQ6**: Are fairness improvements consistent across all protected attribute intersections, or do some subgroups remain persistently disadvantaged?

7.3 Limitations of Aggregate Fairness Metrics

Fairness in AI is commonly assessed using aggregate group-based metrics, as seen in the previous chapters 5 and 6, which compare outcomes across sensitive attributes such as race, gender, or age. These metrics, such as Statistical Parity Difference, Disparate Impact, and Equal Opportunity Difference, offer high-level diagnostic insights and are frequently employed for regulatory compliance. However, they present critical limitations in capturing the complexity of real-world algorithmic harms.

One major shortcoming is their reliance on single-axis evaluations. Most fairness assessments compute disparities along individual attributes, such as race or gender,

but rarely analyze their intersections. As a result, a model may appear fair across racial groups and across genders when evaluated separately, yet still disadvantage subgroups such as Black women or elderly Hispanic men, whose compounded identities are not explicitly measured (Kearns et al. 2018; Ghosh, Genuit, and Reagan 2021; Foulds et al. 2020).

Additionally, aggregate metrics often mask internal variability within demographic groups. A model that performs well on average for women may still underperform significantly for subgroups of women of color or women in older age brackets. This internal heterogeneity is critical to understanding how bias operates within intersectionally defined populations (Buolamwini and Gebru 2018; Hanna et al. 2020a; Mehrabi et al. 2021).

While aggregate metrics remain useful for broad assessments, they are insufficient on their own to ensure fairness across diverse and marginalized communities. To uncover these hidden disparities, fairness evaluations must be disaggregated to the subgroup level. The next section introduces a methodological framework for conducting such subgroup audits, offering a more nuanced and inclusive approach to fairness evaluation.

7.4 Subgroup Auditing Framework

To overcome the limitations of aggregate fairness metrics outlined in Section 7.3, we introduce a subgroup auditing framework that systematically evaluates AI model performance across intersectionally defined groups. This approach builds on the methodological foundation established in Chapter 4, where fairness-enhancing techniques were integrated across the pre-processing, in-processing, and post-processing stages of the machine learning pipeline.

In this chapter, we extend that framework by computing fairness metrics not only at the aggregate level but also across disaggregated subgroups defined by combinations of protected attributes. For example, in the COMPAS and Adult Income datasets, we disaggregate outcomes across race–gender intersections. In our

health insurance dataset, we further incorporate age to examine race–gender–age subgroups.

This granular analysis allows us to detect fairness violations that remain hidden under traditional evaluation paradigms. By integrating fairness-enhancing techniques with subgroup-sensitive metrics, we demonstrate how existing tools can be adapted to reveal intersectional disparities and mitigate compound bias.

Our empirical strategy assesses both baseline and fairness-improved models under this framework, evaluating whether interventions that perform well at the aggregate level also improve outcomes for the most marginalized. In doing so, we support the development of AI systems that are not only technically sound but also aligned with principles of social justice and equity.

7.4.1 Defining Intersectional Subgroups

In the main body of this work, intersectional subgroups are defined using the joint distribution of *race* × *gender*. This design captures the compounded disadvantage experienced by individuals at the intersection of multiple marginalized identities. For example, while "all women" or "all Black individuals" may appear relatively well-treated under aggregate evaluations, the subgroup "Black women" often reveals hidden harms. Focusing on race–gender combinations ensures that disparities affecting doubly marginalized groups are explicitly examined.

It is important to clarify that while a health insurance dataset is also used in this chapter 7.1 (see Section 7.9), the health insurance study serves primarily as a *validation exercise* and a real-world application to demonstrate generalizability of fairness-enhancing integrations. The audit carried out on the health insurance dataset incorporates race, gender, and age intersections, but the core intersectional analysis presented here is based on the two primary datasets, COMPAS and Adult Income, using race–gender subgroup definitions.

7.4.2 Metrics for Subgroup Auditing

The subgroup auditing framework evaluates the same fairness and performance metrics outlined in Chapter 4, section 2.6.1, but disaggregated across race–gender subgroups. For each subgroup, we computed: Statistical Parity Difference (SPD), Disparate Impact (DI), Equal Opportunity Difference (EOD), Predictive Equality Difference (PED), Accuracy (Acc).

This approach enables the identification of fairness gerrymandering, where models may appear fair at the aggregate level (e.g., across all women and all Black individuals separately) while still producing inequitable outcomes for subgroups such as Black women.

7.4.3 Fairness Tools and Extensions

To operationalize subgroup auditing, we extended baseline fairness-enhancing methods with custom configurations suited for intersectional analysis. In the course of our research, we identified several limitations in existing techniques, particularly the Disparate Impact Remover and Adversarial Debiasing methods provided by AIF360 (Bellamy et al. 2019). Although the Disparate Impact Remover supports intersectional attributes, this functionality requires manual configuration, limiting its scalability. In contrast, the Adversarial Debiasing algorithm does not support intersectionality at all in its current implementation.

To address these gaps, we introduced enhanced variants: `DIR+`, an improved version of the Disparate Impact Remover with better handling of small and imbalanced subgroups; and `AD+`, a modified adversarial debiasing framework calibrated for intersectional subgroup constraints. These extensions were applied during the preprocessing and in-processing stages, and their outcomes were evaluated using both aggregate and disaggregated fairness metrics. The use of such tailored tools ensured that fairness-enhancing interventions were sensitive not only to overall disparities, but also to the compounded harms experienced by intersectional minorities.

7.5 Improved Fairness Algorithms: DIR+ and AD+

7.5.1 DIR+: Preprocessing for Intersectional Fairness with Controlled Feature Adjustment

The original *Disparate Impact Remover* (DIR) (Feldman et al. 2015; Bellamy et al. 2019) reduces correlations between protected attributes and other features by applying a repair transformation. However, its baseline design is limited to a single binary attribute, offers no control over fragile subgroups, and lacks transparency about which groups are altered.

We propose **DIR+**, an enhanced preprocessing method tailored for intersectional fairness auditing. DIR+ extends the baseline DIR with intersectional subgroup support, subgroup filtering, selective repair, and robust diagnostics.

Algorithmic Description. Let the dataset be $D = (X, y, G)$, where $X \in \mathbb{R}^{n \times d}$ are features, y are labels, and $G = (g_1, g_2, \ldots, g_k)$ are protected attributes (e.g., race, gender).

DIR+ operates as follows:

1. **Intersectional grouping.** Construct joint subgroup labels:

$$g^{(i)} = \Big|_{j=1}^{k} g_j^{(i)},$$

   where each sample i is assigned an intersectional identity (e.g., `race=Black|gender=Female`).

2. **Group filtering.** For each subgroup g, compute its size $|g|$.

   - If $|g| <$ `min_group_size`, exclude g.
   - If `groups_to_repair` is specified, retain only those groups.

   This avoids overfitting to statistically fragile subgroups.

3. **Repair transformation.** For valid groups, append group codes to the feature matrix and apply the GeneralRepairer (Feldman et al. 2015) with strength

$\lambda \in [0, 1]$:

$$X' = R_\lambda(X, g),$$

where $\lambda = 0$ leaves features unchanged and $\lambda = 1$ enforces maximal decorrelation. Only non-protected features are altered; protected attributes themselves are restored after repair.

4. **Output.** Return the repaired dataset $D' = (X', y, G)$.

Enhancements over DIR. DIR+ introduces the following improvements:

- **Intersectionality support** : repairs correlations across combinations of multiple protected attributes (e.g., race × gender).
- **Selective repair** : applies transformations only to specified subgroups (`groups_to_repair`).
- **Group filtering** : excludes subgroups with insufficient size, controlled by `min_group_size`, to prevent unreliable or overcorrected mappings.
- **Global repair consistency** : ensures a uniform transformation across valid groups for comparability.
- **Verbose logging** : outputs group sizes, inclusion/exclusion criteria, and applied repairs to support transparency and ethical auditing.

Ethical Considerations. The introduction of `min_group_size` directly addresses the statistical instability of learning from small intersectional subgroups, a concern frequently raised in intersectional fairness research. By avoiding overcorrection when data are too sparse, DIR+ balances fairness interventions with reliability, reducing the risk of misleading or ethically questionable adjustments.

The enhanced DIR$^+$ is publicly available at: DIR$^+$

7.5.2 AD+: Enhanced Adversarial Debiasing

Adversarial Debiasing (AD) (B. H. Zhang, Lemoine, and Mitchell 2018) mitigates unfairness by jointly training a predictor and an adversary. The predictor f_θ learns to predict labels $\hat{y}$, while the adversary h_ϕ attempts to infer the protected attribute a from the predictor's outputs. The predictor is penalized when the adversary succeeds, thereby discouraging the encoding of group information in predictions.

We propose **AD+**, an enhanced implementation that improves training stability, supports multiple subgroup values, and introduces a more robust gradient manipulation strategy.

Algorithmic Description. Let the dataset be $D = (X, y, a)$, where $X \in \mathbb{R}^{n \times d}$ are features, $y \in \{0, 1\}$ are binary labels, and a is a protected attribute taking values in $\{1, \ldots, m\}$.

AD+ proceeds as follows:

1. **Classifier.** A neural classifier f_θ maps x to prediction $\hat{y}$ via logits s:

$$s = f_\theta(x), \quad \hat{y} = \sigma(s),$$

trained with binary cross-entropy loss:

$$\mathcal{L}_{\text{clf}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell_{\text{BCE}}(y_i, \hat{y}_i).$$

2. **Adversary.** The adversary h_ϕ receives the classifier's probabilistic outputs and their interactions with labels:

$$u = \big[\sigma(s),\, \sigma(s) \cdot y,\, \sigma(s) \cdot (1 - y)\big],$$

and predicts the protected attribute distribution $\hat{a} = h_\phi(u)$ using softmax. Its loss is

$$\mathcal{L}_{\text{adv}}(\phi) = \frac{1}{n} \sum_{i=1}^{n} \ell_{\text{CE}}(a_i, \hat{a}_i).$$

3. **Joint optimization with gradient manipulation.** The classifier's gradients are updated using a two-step modification:

$$\text{Projection:} \quad g_{\text{proj}} = g - (g \cdot \hat{g}_{\text{adv}})\hat{g}_{\text{adv}}$$
$$\text{Adversarial subtraction:} \quad g_{\text{final}} = g_{\text{proj}} - \lambda \cdot g_{\text{adv}}$$

   where $g = \nabla_\theta \mathcal{L}_{\text{clf}}$, $g_{\text{adv}} = \nabla_\theta \mathcal{L}_{\text{adv}}$, $\hat{g}_{\text{adv}}$ is the unit vector of g_{adv}, and λ is the adversarial loss weight. This removes the adversary-aligned gradient component and applies direct adversarial pressure.

4. **Stabilization techniques.** AD+ incorporates gradient clipping (to $[-5, 5]$), dropout, learning-rate decay for the Adam optimizer, and input validation (e.g., NaN/Inf checks) to improve training stability.

Enhancements over AD. AD+ introduces several key improvements over the baseline adversarial debiasing:

- **Multi-class adversary** : supports protected attributes with more than two values using softmax cross-entropy.
- **Combined gradient strategy** : applies both projection and direct subtraction for more effective bias removal.
- **Input enrichment** : adversary receives sigmoid-transformed logits and their interaction with labels.
- **Enhanced stabilization** : gradient clipping, learning-rate scheduling, dropout, and explicit input validation.
- **Explicit control** : adversarial strength is tunable via λ (the `adversary_loss_weight`).

Key Implementation Detail. Unlike approaches that use only projection or only direct subtraction, AD+ employs a **two-step gradient modification**: first pro-

jecting out the adversary-aligned component, then subtracting a weighted adversary gradient. This provides stronger fairness enforcement while maintaining stability.

Ethical Considerations. By extending adversarial debiasing to multi-valued protected attributes and providing more stable training, AD+ makes it feasible to apply this technique in intersectional settings. The input validation and gradient stabilization reduce the risk of training collapse, making fairness enforcement more reliable.

The source code for AD$^+$ is publicly available at: AD$^+$

7.5.3 Ethical Considerations and Design Intent

Both `DIR+` and `AD+` were developed to advance fairness interventions in datasets where group identities are intersectional and multi-valued. By incorporating selective subgroup repair, group filtering, and multi-class adversarial learning, these methods promote proportional fairness, mitigating harm where sufficient data support exists and avoiding overreach in data-scarce contexts.

We evaluated the enhanced methods alongside their original counterparts and found that `DIR+` and `AD+` perform comparably in binary group settings, while significantly extending their applicability to complex, real-world intersectional scenarios. Empirical results demonstrate that these approaches maintain parity with their predecessors in traditional fairness tasks, yet offer greater flexibility and robustness when addressing overlapping or multi-dimensional protected attributes.

Both enhanced methods are fully integrated into our broader fairness auditing pipeline and have been released for public use to support ongoing research in intersectional, context-sensitive, and accountable algorithmic systems.

7.6 Datasets and Experimental Setup

The experiments in this chapter build upon the datasets already introduced in Chapter 4, namely the COMPAS dataset (Section 4.4.1), the Adult Income dataset (Section 4.4.3), and the Irish health insurance dataset (Section 4.4.4). Detailed descrip-

tions of the variables, preprocessing steps, and baseline distributions are provided in those sections; here, we briefly summarize their role in the intersectional auditing framework.

7.6.1 Datasets for Intersectional Auditing

The core intersectional analysis in this chapter draws on the COMPAS and Adult Income datasets, where subgroups are defined as the cross-product of *race* and *gender*. This allows us to surface compounded disparities faced by doubly marginalized groups (e.g., Black women). These datasets provide complementary evaluation contexts: criminal justice (COMPAS) and economic opportunity (Adult Income).

To assess the generalizability of the fairness pipeline, we further include the Irish health insurance dataset (Section 4.4.4) as a validation case. Unlike COMPAS and Adult Income, this dataset adopts an extended subgroup definition based on *race* $\times$ *gender* $\times$ *age*, yielding ten demographic subgroups. Its inclusion demonstrates the applicability of the auditing framework in a high-stakes health insurance setting beyond standard benchmark datasets.

7.6.2 Experimental Setup

All models were trained and evaluated under the multi-phase fairness pipeline introduced in Chapter 4. Fairness-enhancing techniques were applied at pre-, in-, and post-processing stages, both individually and in combination. For subgroup auditing, the metrics Statistical Parity Difference (SPD), Disparate Impact (DI), Equal Opportunity Difference (EOD), Predictive Equality Difference (PED), and predictive Accuracy were computed separately for each subgroup, in addition to the aggregate scores. This setup ensures that hidden disparities, which may not be visible in global evaluations, are surfaced in the intersectional analysis.

7.6.3 Intersectional Analysis of COMPAS and Adult Income Datasets

To jointly evaluate fairness across both the COMPAS and Adult Income datasets, we performed intersectional subgroup analysis using a shared encoding scheme based on combinations of race and gender. As shown in Table 7.1, the reference group (Group 0) corresponds to White males in the Adult dataset and Caucasian males in COMPAS. Other subgroup values represent intersections of race and gender, allowing consistent comparison across datasets.

Table 7.2 presents the aggregate results for several fairness-enhancing configurations. Each model is evaluated using standard group fairness metrics, Statistical Parity Difference (SPD), Disparate Impact (DI), Equal Opportunity Difference (EOD), and Predictive Equality Difference (PED), alongside classification accuracy. These results provide a high-level view of how mitigation strategies influence overall group-level equity and predictive performance.

To reveal deeper intersectional disparities, we conducted disaggregated evaluations by subgroup, as shown in Table 7.3 for individual techniques, Table 7.4 for two-stage techniques, and Table 7.5 for three-stage techniques. This subgroup-level breakdown highlights how different mitigation techniques affect specific demographic intersections, such as Black women or White women, within each dataset. In doing so, we assess not just whether models reduce average bias, but whether they meaningfully improve outcomes for historically marginalized intersections.

The following sections analyze these results in detail, highlighting trade-offs between fairness and accuracy, the effectiveness of mitigation strategies across contexts, and the persistence of disparities that aggregate metrics often overlook.

Table 7.1: Encoded representation of intersectional subgroups in the Adult and COMPAS datasets. Group 0 represents the privileged reference group (Male White for Adult; Male Caucasian for COMPAS), with other groups reflecting intersections of race and gender.

Adult Group		**COMPAS Group**	
Value	Representation	Value	Representation
Male_White	0	Male_Caucasian	0
Female_White	1	Female_Caucasian	1
Male_Black	2	Male_African-American	2
Female_Black	3	Female_African-American	3

Table 7.2: Aggregate fairness and accuracy metrics for each model configuration on the Adult and COMPAS datasets.

Model	Adult					COMPAS				
	SPD	DI	EOD	PED	Acc	SPD	DI	EOD	PED	Acc
NONE	-0.3248	0.3206	-0.1761	-0.2179	0.81	0.2054	1.7159	0.2143	0.1404	0.68
Re	-0.1990	0.5146	0.0227	-0.1018	0.82	0.1614	1.4943	0.1410	0.1198	0.67
DIR	-0.2660	0.4373	-0.0632	-0.1677	0.80	0.2065	1.6577	0.2037	0.1484	0.68
EGR	-0.1905	0.5319	0.0321	-0.0933	0.81	-0.0428	0.9160	-0.0498	-0.0089	0.61
AL	-0.0642	0.8146	0.1392	0.0336	0.80	0.1426	1.3649	0.1028	0.1352	0.65
EO	-0.1125	0.7256	-0.0173	0.0022	0.78	0.0294	1.0761	0.0036	-0.0003	0.65
ROC	-0.1990	0.5146	0.0227	-0.1018	0.82	0.1614	1.4943	0.1410	0.1198	0.67
Re+EGR	-0.1907	0.5314	0.0338	-0.0938	0.82	-0.0718	0.8665	-0.0795	-0.0337	0.62
Re+AL	0.0642	1.2276	-0.1392	-0.0336	0.80	0.1772	1.5044	0.1615	0.1359	0.66
DIR+EGR	0.2285	1.9709	0.0246	0.1383	0.78	-0.1265	0.6779	-0.1814	-0.0600	0.58
DIR+AL	-0.1218	0.6663	-0.0238	-0.0093	0.80	0.1416	1.2338	0.0431	0.1853	0.62
Re+EO	-0.1125	0.7256	-0.0173	0.0022	0.78	0.0294	1.0761	0.0036	-0.0003	0.65
Re+ROC	-0.1990	0.5146	0.0227	-0.1018	0.82	0.1614	1.4943	0.1410	0.1198	0.67
DIR+EO	-0.1078	0.7646	-0.0022	-0.0004	0.74	0.0283	1.0757	0.0033	-0.0047	0.65
DIR+ROC	-0.2660	0.4373	-0.0632	-0.1677	0.80	0.2065	1.6577	0.2037	0.1484	0.68
EGR+EO	-0.1127	0.7231	-0.0061	0.0007	0.78	0.0137	1.0294	0.0005	-0.0004	0.59
EGR+ROC	-0.1942	0.5243	0.0274	-0.0970	0.82	-0.1681	0.7720	-0.0556	-0.2866	0.59
AL+EO	-0.1104	0.6825	0.0200	-0.0025	0.81	0.0207	1.0427	0.0080	-0.0060	0.64
AL+ROC	-0.0980	0.7746	0.0721	0.0088	0.77	0.0633	1.0852	-0.0063	0.0891	0.59
Re+EGR+EO	-0.1127	0.7236	-0.0026	0.0004	0.78	0.0156	1.0334	0.0038	-0.0034	0.61
Re+EGR+ROC	-0.1942	0.5243	0.0274	-0.0970	0.82	-0.0180	0.9469	-0.0288	-0.0493	0.66
Re+AL+EO	-0.1104	0.6825	0.0200	-0.0025	0.81	0.0281	1.0682	0.0080	-0.0043	0.64
Re+AL+ROC	-0.0980	0.7746	0.0721	0.0088	0.77	0.0623	1.0922	0.0470	0.0319	0.61
DIR+EGR+EO	-0.1040	0.7757	0.0049	-0.0007	0.74	0.0059	1.0221	0.0129	-0.0126	0.55
DIR+EGR+ROC	-0.1768	0.6209	0.0060	-0.0776	0.77	0.0177	1.3875	0.0061	0.0153	0.56
DIR+AL+EO	-0.1076	0.6963	-0.0019	0.0008	0.80	0.0158	1.0216	-0.0214	0.0153	0.60
DIR+AL+ROC	-0.1466	0.6695	-0.0318	-0.0364	0.77	0.0153	1.0159	-0.0114	0.0314	0.49

Table 7.3: Intersectional subgroup-level evaluation of fairness and accuracy metrics for individual fairness mitigation techniques on the Adult and COMPAS datasets.

Model	Group	Adult					COMPAS				
		SPD	DI	EOD	PED	SAcc	SPD	DI	EOD	PED	SAcc
NONE	1	-0.3153	0.3403	-0.1295	-0.2141	0.90	-0.1266	0.5586	-0.1881	-0.0736	0.67
	2	-0.2857	0.4023	-0.2403	-0.2017	0.86	0.2999	2.0454	0.2906	0.2179	0.68
	3	-0.4107	0.1409	-0.4705	-0.2497	0.92	0.0612	1.2133	0.0440	0.0764	0.65
Re	1	-0.2016	0.5081	0.0260	-0.1058	0.87	0.1351	1.4140	0.1835	0.1286	0.68
	2	-0.1422	0.6530	0.0198	-0.0841	0.85	0.1666	1.5105	0.1299	0.1085	0.66
	3	-0.2408	0.4127	-0.0037	-0.0982	0.87	0.1609	1.4931	0.1806	0.1528	0.66
DIR	1	-0.2929	0.3803	-0.0969	-0.1941	0.89	0.0322	1.1026	0.0383	0.0490	0.68
	2	-0.1004	0.7877	0.0625	-0.0435	0.78	0.2602	1.8289	0.2376	0.1865	0.68
	3	-0.2900	0.3864	-0.0869	-0.1472	0.85	0.1101	1.3508	0.1155	0.1111	0.66
EGR	1	-0.1980	0.5133	0.0261	-0.1015	0.87	-0.0067	0.9856	0.00001	0.0140	0.59
	2	-0.1069	0.7372	0.0559	-0.0488	0.83	0.0633	1.1357	0.0431	0.0314	0.60
	3	-0.2333	0.4267	0.0238	-0.0918	0.87	-0.0116	0.9751	0.1541	-0.0696	0.69
AL	1	-0.1161	0.6647	0.1077	-0.0216	0.85	0.2377	1.6083	0.2041	0.2956	0.55
	2	0.1031	1.2978	0.2156	0.1710	0.72	0.1060	1.2713	0.0664	0.0575	0.66
	3	0.0371	1.1072	0.2410	0.1794	0.67	0.2317	1.5955	0.3109	0.2351	0.63
EO	1	-0.1178	0.7127	-0.0147	-0.0045	0.78	-0.0213	0.9449	0.0706	-0.0543	0.72
	2	-0.0653	0.8407	-0.0222	0.0195	0.76	0.0328	1.0849	-0.0152	-0.0011	0.62
	3	-0.1315	0.6792	-0.0287	0.0196	0.75	0.0627	1.1621	0.0746	0.0590	0.65
ROC	1	-0.2016	0.5081	0.0260	-0.1058	0.87	0.1351	1.4140	0.1835	0.1286	0.68
	2	-0.1422	0.6530	0.0198	-0.0841	0.85	0.1666	1.5105	0.1299	0.1085	0.66
	3	-0.2408	0.4127	-0.0037	-0.0982	0.87	0.1609	1.4931	0.1806	0.1528	0.66

Table 7.4: Intersectional subgroup–level fairness and accuracy for two-stage combinations of mitigation methods across PI, IP, and PP on the Adult and COMPAS datasets.

Model	Group	Adult					COMPAS				
		SPD	DI	EOD	PED	SAcc	SPD	DI	EOD	PED	SAcc
Re+EGR	1	-0.1989	0.5112	0.0261	-0.1025	0.87	0.0068	1.0146	-0.0138	0.0442	0.57
	2	-0.1039	0.7448	0.0643	-0.0469	0.83	0.1038	1.2226	0.0923	0.0532	0.63
	3	-0.2333	0.4267	0.0238	-0.0918	0.87	-0.0246	0.9472	0.0618	-0.0402	0.66
DIR+EGR	1	-0.2328	0.4980	-0.0347	-0.1423	0.84	0.0223	1.0840	0.0566	0.0069	0.61
	2	-0.1499	0.6767	0.0119	-0.0987	0.81	0.1572	1.5908	0.1926	0.0950	0.57
	3	-0.2827	0.3906	-0.0285	-0.1528	0.85	0.0757	1.2843	0.2156	-0.0139	0.67
Re+AL	1	-0.1161	0.6647	0.1077	-0.0216	0.85	0.3281	1.9339	0.3338	0.3439	0.58
	2	0.1031	1.2978	0.2156	0.1710	0.72	0.1303	1.3708	0.1241	0.0428	0.68
	3	0.0371	1.1072	0.2410	0.1794	0.67	0.626	1.7473	0.2670	0.2639	0.61
DIR+AL	1	-0.1879	0.4853	-0.0911	-0.0756	0.86	0.1581	1.2611	0.0461	0.2605	0.47
	2	0.1043	1.2857	0.1546	0.1861	0.68	0.1188	1.1962	0.0324	0.1160	0.66
	3	-0.0057	0.9844	0.1470	0.1400	0.69	0.2387	1.3943	0.1301	0.3395	0.45
Re+EO	1	-0.1178	0.7127	-0.0147	-0.0045	0.78	-0.0213	0.9449	0.0706	-0.0543	0.72
	2	-0.0653	0.8407	-0.0222	0.0195	0.76	0.038	1.0849	-0.0152	-0.0011	0.62
	3	-0.1315	0.6792	-0.0287	0.0196	0.75	0.0627	1.1621	0.0746	0.0590	0.65
Re+ROC	1	-0.2016	0.5081	0.0260	-0.1058	0.87	0.1351	1.4140	0.1835	0.1286	0.68
	2	-0.1422	0.6530	0.0198	-0.0841	0.85	0.1666	1.5105	0.1299	0.1085	0.66
	3	-0.2408	0.4127	-0.0037	-0.0982	0.87	0.1609	1.4931	0.1806	0.1528	0.66
DIR+EO	1	-0.1362	0.7027	-0.0329	-0.0288	0.75	-0.1306	0.6509	-0.1921	-0.0769	0.65
	2	0.0481	1.1050	0.1074	0.1250	0.66	0.0744	1.1989	0.0383	0.0219	0.65
	3	-0.1152	0.7484	-0.0088	0.0285	0.70	-0.0451	0.8795	-0.0654	-0.0278	0.66
DIR+ROC	1	-0.2929	0.3803	-0.0969	-0.1941	0.89	0.0322	1.1026	0.0383	0.0490	0.68
	2	-0.1004	0.7877	0.0625	-0.0435	0.78	0.2602	1.8289	0.2376	0.1865	0.68
	3	-0.2900	0.3864	-0.0869	-0.1472	0.85	0.1101	1.3508	0.1155	0.1111	0.66
EGR+EO	1	-0.1160	0.7150	-0.0099	-0.0029	0.78	-0.0067	0.9856	-0.0204	0.0241	0.57
	2	-0.0315	0.9225	0.0139	0.0529	0.74	0.0254	1.0544	-0.0107	0.0136	0.58
	3	-0.1749	0.5702	-0.0262	-0.0265	0.80	-0.0246	0.9472	0.1148	-0.0696	0.68

Table 7.4: Intersectional subgroup–level fairness and accuracy for two-stage combinations of mitigation methods across PI, IP, and PP on the Adult and COMPAS datasets.

Model	Group	Adult					COMPAS				
		SPD	DI	EOD	PED	SAcc	SPD	DI	EOD	PED	SAcc
EGR+ROC	1	-0.2004	0.5094	0.0249	-0.1041	0.87	-0.2302	0.6876	-0.1165	-0.2790	0.59
	2	-0.1161	0.7158	0.0463	-0.0578	0.83	-0.1422	0.8071	-0.0570	-0.2682	0.64
	3	-0.239	0.4142	-0.0025	-0.0965	0.87	-0.2369	0.6784	0.0147	-0.3541	0.69
AL+EO	1	-0.1566	0.5493	-0.0134	-0.0510	0.86	0.1178	1.2436	0.1396	0.1418	0.58
	2	0.0401	1.1154	0.0883	0.1224	0.73	-0.0117	0.9759	-0.0261	-0.0766	0.65
	3	-0.0212	0.9389	0.1647	0.1222	0.72	0.0879	1.1817	0.1680	0.0830	0.63
AL+ROC	1	-0.1590	0.6342	0.0466	-0.0576	0.82	0.0546	1.0734	-0.0099	0.1241	0.47
	2	0.0792	1.1822	0.1389	0.1621	0.66	0.0403	1.0542	-0.0153	0.0206	0.66
	3	0.0399	1.0918	0.1391	0.1943	0.58	0.1858	1.2502	0.0726	0.2796	0.39

Table 7.5: Intersectional subgroup–level fairness and accuracy for three-stage combinations of mitigation methods across PIP on the Adult and COMPAS datasets.

Model	Group	Adult					COMPAS				
		SPD	DI	EOD	PED	SAcc	SPD	DI	EOD	PED	SAcc
Re+EGR+EO	1	-0.1201	0.7049	-0.0099	-0.0076	0.78	-0.0067	0.9856	-0.0342	0.0341	0.57
	2	-0.0254	0.9376	0.02223	0.0586	0.74	0.0345	1.0740	0.0129	-0.0031	0.61
	3	-0.1584	0.6107	-0.0012	-0.0105	0.79	-0.0571	0.8776	-0.0362	-0.0402	0.62
Re+EGR+ROC	1	-0.2004	0.5094	0.0249	-0.1041	0.87	-0.3385	0.00001	-0.5023	-0.2200	0.67
	2	-0.1161	0.7158	0.0463	-0.0578	0.84	0.1086	1.3207	0.0818	0.0530	0.65
	3	-0.2392	0.4142	-0.0025	-0.0965	0.87	-0.3385	0.00001	-0.5023	-0.2200	0.67
Re+AL+ROC	1	-0.1590	0.6342	0.0466	-0.0576	0.82	0.1320	1.1954	0.1508	0.1367	0.53
	2	0.0792	1.1822	0.1389	0.1621	0.66	0.0125	1.0186	0.0183	-0.0707	0.68
	3	0.0399	1.0918	0.1391	0.1943	0.58	0.2420	1.3582	0.1543	0.3021	0.45
Re+AL+EO	1	-0.1566	0.5493	-0.0134	-0.0510	0.86	0.1461	1.3548	0.1979	0.1344	0.63
	2	0.0401	1.1154	0.0883	0.1224	0.73	-0.0061	0.9853	-0.0285	-0.0661	0.64
	3	-0.0212	0.9389	0.1647	0.1222	0.72	0.080	1.1993	0.0913	0.0799	0.63
DIR+EGR+ROC	1	-0.1693	0.6371	-0.0010	-0.0684	0.79	0.0825	2.8030	0.162	0.0626	0.67
	2	-0.1110	0.7620	0.0493	-0.0549	0.79	-0.0166	0.6373	-0.0420	-0.0089	0.47
	3	-0.2793	0.4012	-0.0499	-0.1403	0.85	0.1251	3.7362	0.2345	0.0556	0.69
DIR+EGR+EO	1	-0.1060	0.7714	-0.0011	-0.0025	0.72	-0.0289	0.8913	-0.0862	0.0069	0.56
	2	-0.0376	0.9188	0.0203	0.0369	0.70	0.0165	1.061	0.0336	-0.0233	0.53
	3	-0.1584	0.6585	0.0215	-0.0238	0.73	-0.0129	0.9513	-0.0425	0.0069	0.56
DIR+AL+ROC	1	-0.2121	0.5219	-0.0744	-0.1054	0.83	0.0387	1.0402	0.00001	0.0667	0.33
	2	0.0594	1.1339	0.0754	0.1491	0.65	0.0060	1.0062	-0.0140	0.0103	0.58
	3	-0.0140	0.9684	0.0928	0.1347	0.62	0.0387	1.0402	0.00001	0.0667	0.33
DIR+AL+EO	1	-0.1738	0.5093	-0.0685	-0.0657	0.86	0.0324	1.0443	-0.0184	0.0905	0.47
	2	0.1167	1.3294	0.1748	0.1945	0.68	-0.0070	0.9905	-0.0321	-0.0540	0.66
	3	0.0112	1.0315	0.1672	0.1528	0.68	0.1130	1.1546	0.0656	0.1695	0.45

7.7 Results and Analysis

This section presents the key findings in relation to the central hypothesis and the guiding research questions.

Integrated fairness-enhancing techniques that incorporate intersectional protected attributes, such as race and gender jointly, can achieve improved fairness outcomes while maintaining competitive predictive accuracy. This hypothesis is grounded in the idea that mitigating compound forms of discrimination through targeted fairness interventions and intersectional evaluation enables ML models to address bias more comprehensively. This becomes especially relevant when multiple mitigation strategies are combined, as their cumulative effect may better capture the nuanced patterns of bias that affect marginalized subgroups.

Table 7.1 outlines the intersectional subgroup encodings applied to the Adult and COMPAS datasets. In both cases, Group 0 represents the privileged reference category, White males in the Adult dataset and Caucasian males in COMPAS. The remaining groups correspond to unique combinations of race and gender.

RQ1: Do integrated fairness-enhancing models reduce disparities across intersectional subgroups (race × gender), and do aggregate metrics obscure these disparities?

The results show that while integrated fairness-aware models can substantially reduce disparities among intersectional subgroups, aggregate-level metrics often conceal lingering inequities. In the COMPAS dataset, for instance, the unmitigated model (`NONE`) reported an overall Statistical Parity Difference (SPD) of 0.2054, suggesting a moderate level of bias. Yet, when disaggregated by subgroup, more severe disparities emerged: Male African-Americans exhibited an SPD of 0.2999, which is approximately 45% higher than the dataset-wide figure. Similarly, Female African-Americans had lower subgroup accuracy (0.65) compared to the privileged group (0.68).

The Adult dataset revealed parallel patterns. An aggregate Disparate Impact (DI) of 0.3206 failed to reflect the degree of marginalization faced by Female Black

individuals, who experienced a subgroup DI as low as 0.1409, well below fairness thresholds.

Mitigation techniques such as the combined strategy `Re+EGR+EO` resulted in marked improvements. For example, the SPD for Female Black individuals in the Adult dataset improved from -0.4107 to 0.1584. Despite these gains, true parity was not reached, particularly for multiply marginalized subgroups. These outcomes underscore the inadequacy of aggregate fairness metrics in capturing the lived realities of algorithmic harm. Instead, intersectional analysis is essential to surface compounded discrimination, especially for groups such as Black women who repeatedly experienced the most adverse outcomes. This aligns with the concept of *fairness gerrymandering*, where models may appear fair on average while disproportionately disadvantaging minority subgroups.

RQ2: What trade-offs arise between fairness metrics (e.g., SPD, DI, EOD) and predictive accuracy when incorporating intersectionality?

The findings reveal a complex landscape of trade-offs between fairness objectives and predictive performance, which differ significantly across datasets and subgroups. In the COMPAS dataset, for example, post-processing techniques such as `Equalized Odds (EO)` were able to achieve near-perfect fairness with respect to Equal Opportunity Difference (EOD = 0.0036). However, this improvement came at a cost: the overall model accuracy dropped to 0.65 from the baseline of 0.68. More intensive mitigation configurations, such as `DIR+AL+ROC`, led to even more pronounced trade-offs, reducing the subgroup accuracy for Female African-Americans to just 0.33, despite measurable gains in Disparate Impact (DI).

By contrast, the Adult dataset exhibited more moderate trade-offs. The integrated configuration `Re+EGR+EO` achieved a 61% improvement in Statistical Parity Difference (SPD) while maintaining a relatively high accuracy of 0.78, only slightly below the baseline of 0.81. Nonetheless, the pursuit of fairness across one metric often undermined performance in others. For instance, although `Adversarial Learning (AL)` improved the DI score for Male Black adults (1.2978), it introduced

a reverse disparity in SPD (+0.1031), favoring this group disproportionately.

These outcomes highlight a key normative challenge in fairness optimization: enhancing one fairness dimension may inadvertently compromise another. Achieving balance across metrics such as SPD, DI, EOD, and PED is often infeasible within a single model. As such, practitioners must make deliberate, context-sensitive decisions about which fairness criteria to prioritize, whether that means minimizing false positives in criminal justice settings or addressing representation gaps in employment contexts. The study affirms that fairness is inherently multidimensional and must be navigated through thoughtful, domain-aware trade-offs.

RQ3: Which fairness mitigation strategies (individual or integrated) are most effective for improving subgroup fairness while maintaining accuracy?

The results suggest that integrated, multi-stage mitigation approaches are generally more effective than individual strategies in balancing fairness and predictive accuracy. In the Adult dataset, the combined method `Re+EGR+EO` demonstrated strong overall performance, improving Statistical Parity Difference (SPD) from -0.3248 to -0.1127, while maintaining a relatively high accuracy of 0.78, down only slightly from the baseline of 0.81. Importantly, subgroup accuracy remained consistently high, with values above 0.74, indicating that the model's improvements in fairness did not come at the cost of severe performance degradation for any group.

In contrast, the COMPAS dataset revealed a slightly different pattern. The most effective configuration was not an integrated approach, but the post-processing method `Equalized Odds (EO)`, which achieved excellent fairness metrics (EOD = 0.0036, SPD = 0.0294) while keeping accuracy at a reasonable level (0.65). This suggests that the effectiveness of mitigation strategies is context-dependent and can vary significantly across domains.

When examined individually, the mitigation techniques showed notable limitations. Pre-processing methods like `Reweighing` were effective in the Adult dataset, where representational imbalance was a core issue, but had limited impact in the

more error-sensitive COMPAS context. In-processing techniques, such as `Adversarial Learning (AL)`, introduced high variability, sometimes improving fairness for one metric while exacerbating others. Interestingly, the hybrid strategy `Re+EGR+ROC` performed well in COMPAS, balancing fairness (SPD = -0.0180) with a moderate accuracy of 0.66.

These findings reinforce the idea that algorithmic bias is complex and multi-dimensional. No single intervention suffices; instead, a layered approach is often required, combining pre-processing to adjust for structural imbalances, in-processing to enforce fairness during learning, and post-processing to correct residual disparities. The optimal configuration must be carefully tailored to the specific characteristics and fairness objectives of the application domain.

RQ4: How do fairness mitigation strategies perform differently across domains (e.g., COMPAS vs. Adult datasets)?

The results reveal clear domain-specific distinctions in how fairness interventions behave across different application contexts. In the Adult dataset, which centers on income prediction, the dominant form of bias arises from underrepresentation. This is evident in the markedly low Disparate Impact (DI) score of 0.1409 for Female Black individuals. In this context, pre-processing techniques such as `Reweighing` proved particularly effective, substantially improving fairness metrics with minimal compromise to predictive accuracy.

By contrast, the COMPAS dataset, which deals with criminal recidivism prediction, exhibits a different fairness challenge: over-prediction bias. For example, Male African-Americans had an SPD of 0.2999, indicating a disproportionately high rate of positive predictions. These kinds of disparities are harder to mitigate through data balancing alone. Instead, post-processing approaches such as `Equalized Odds (EO)` were more effective in reducing these disparities, although they often incurred a larger drop in model accuracy.

These contrasting patterns reflect deeper structural and institutional biases embedded in each domain. The Adult dataset mirrors labor market inequities, where

representational imbalances dominate, whereas COMPAS encodes systemic biases present in the judicial system, resulting in error disparities. As such, fairness strategies must be aligned with the specific characteristics of the domain: rebalancing methods suit representation-driven datasets like Adult, while outcome correction techniques are better suited to domains like COMPAS, where misclassification risk is higher.

RQ5: How does the stage of intervention (pre-, in-, or post-processing) influence fairness outcomes?

The effectiveness of fairness mitigation is strongly influenced by the stage at which the intervention is applied. In the Adult dataset, pre-processing methods such as `Reweighing` were especially effective. For example, this technique improved the Statistical Parity Difference (SPD) for Female Black individuals from -0.4107 to -0.2408, demonstrating its strength in correcting representational imbalances.

In-processing interventions produced more variable results. While `EGR` led to meaningful fairness improvements in Adult, other in-processing methods such as `Adversarial Learning (AL)` introduced unintended effects. Specifically, AL produced a reverse disparity for Male Black individuals, resulting in an SPD of $+0.1031$, an overcorrection that shifted bias in the opposite direction.

In contrast, post-processing was the most effective stage of intervention in the COMPAS dataset. The `Equalized Odds (EO)` method delivered near-perfect fairness in terms of Equal Opportunity Difference (EOD = 0.0036), while maintaining a subgroup accuracy of 0.62, which is acceptable given the complexity of the domain and the observed bias patterns.

These results highlight the importance of a pipeline-aware approach to fairness. Each mitigation stage addresses a different aspect of bias: pre-processing tackles data imbalance, in-processing embeds fairness during model training, and post-processing corrects residual disparities in predictions. Selecting the appropriate stage, or integrating multiple stages, is essential for tailoring fairness strategies to the specific sources and manifestations of bias in a given dataset or domain.

RQ6: Are fairness improvements consistent across all protected attribute intersections, or do some subgroups remain disadvantaged?

The analysis confirms that fairness improvements are not evenly distributed across all subgroups. Intersectionally marginalized populations, those affected by multiple dimensions of disadvantage, often continue to experience disproportionate harms even after mitigation. For instance, in the Adult dataset, the SPD for Female Black individuals improved from -0.4107 to -0.1584 using the integrated strategy `Re+EGR+EO`, yet this remained significantly worse than the reference group (SPD $= 0$), indicating residual inequity.

Similarly, in the COMPAS dataset, the combined approach `DIR+AL+ROC` improved aggregate fairness metrics, but at the cost of severely reduced accuracy for the Female African-American subgroup, dropping to just 0.33. This outcome illustrates that certain mitigation strategies, while effective at the population level, may inadvertently exacerbate disparities for the most vulnerable intersections.

Such inconsistencies reflect the challenges raised by intersectionality theory, which emphasizes the compounded nature of disadvantage at the overlap of multiple identity categories. These findings suggest that models may preferentially benefit subgroups with simpler or more dominant bias patterns, while overlooking those at the intersection of multiple marginalizations.

To address these disparities, it is critical to incorporate disaggregated evaluation frameworks and adopt subgroup-specific mitigation methods, such as targeted reweighing or adversarial debiasing, to ensure that fairness gains extend to those most at risk.

In conclusion, the findings support the central hypothesis: intersectional, multi-stage fairness interventions can enhance equity without significantly compromising model performance. However, these improvements are fragile and uneven, underscoring the need for context-sensitive, subgroup-aware, and ethically informed approaches to fairness evaluation.

7.8 Comparative Analysis of Fairness-Accuracy Trade-offs: Adult vs. COMPAS Datasets

A comparative analysis of the Adult and COMPAS datasets reveals fundamentally different dynamics in how algorithmic bias manifests across domains. While the Adult dataset, centered on income prediction, predominantly reflects the underrepresentation of marginalized groups, the COMPAS dataset, focused on recidivism prediction, exhibits instability in fairness outcomes across intersectional subgroups (see Tables 7.6 and 7.7).

To examine general trends across mitigation strategies, we evaluated aggregate performance using summary statistics, including means, standard deviations, and 95% confidence intervals (CIs) for key fairness metrics and predictive accuracy (Table 7.6). In the Adult dataset, the average Statistical Parity Difference (SPD) was moderately negative (mean = -0.138, CI = $[-0.184, -0.092]$), and the Disparate Impact (DI) was well below the commonly accepted threshold of 0.8 (mean = 0.670), indicating ongoing group-level disparities despite applied fairness interventions. Conversely, the COMPAS dataset displayed a reversal of disparity direction, with an average DI of 1.147 (CI = [1.016, 1.279]). Equal Opportunity Difference (EOD) remained close to zero in both datasets but with wide confidence intervals, suggesting variability rather than consistent improvement. Predictive Equality Difference (PED) showed higher variability in Adult (mean = -0.050, CI = $[-0.082, -0.019]$). In terms of predictive accuracy, Adult models outperformed those in COMPAS, with an average accuracy of 0.793 versus 0.624, respectively. These findings underscore that, while multi-stage mitigation strategies can improve group-level metrics, they do not eliminate variance or guarantee consistent fairness when intersectional subgroups are considered.

The Adult dataset consistently reflects systemic exclusion across fairness metrics (Table 7.7). For example, the SPD for Black women (Group 3) is significantly negative (mean = -0.1506, CI = $[-0.1959, -0.1053]$), indicating that they are

approximately 15% less likely than the privileged group (White men) to receive favorable predictions. DI values for all unprivileged groups are below 1.0, including White women (Group 1) at 0.5599 and Black women at 0.5889, suggesting persistent disadvantage in outcomes. Subgroup accuracy (SAcc) reflects this disparity: the privileged group achieves 83.78% accuracy, while Black men and Black women lag behind at 76.26% and 78.89%, respectively. These gaps point to a fairness-accuracy trade-off where the model sustains overall performance partly by replicating historical inequities.

In contrast, the COMPAS dataset reveals a more complex pattern. Black male defendants (Group 2) exhibit a positive SPD of 0.1088 and a DI of 1.2757, which might suggest favorable treatment. However, this likely reflects over-prediction of recidivism risk, a deeply problematic outcome in the criminal justice context. Once again, Black women (Group 3) suffer the poorest outcomes, with the lowest subgroup accuracy at 60.22%, underscoring the compounded disadvantages experienced at the intersection of race and gender.

Error-related fairness metrics reinforce these insights. In the Adult dataset, Black men show elevated rates for both true positives (EOD = 0.0684) and false positives (PED = 0.0448), indicating uneven decision thresholds. In COMPAS, fairness metrics such as EOD and PED show wider confidence intervals, reflecting greater variability and reduced stability across demographic subgroups. This may be attributed to the inherent difficulty of predicting complex human behaviors in judicial settings, in contrast to the relatively structured nature of income prediction.

These results have important implications for fairness research and applied AI. In the Adult dataset, interventions should focus on correcting entrenched underrepresentation while preserving accuracy. Approaches such as reweighing, data augmentation, or fairness-aware optimization are particularly relevant. For the COMPAS dataset, however, the priority should shift toward achieving consistency and stability in outcomes across subgroups, even at some cost to overall accuracy.

Across both datasets, the persistent disadvantage faced by Black women high-

lights the critical need for intersectional analysis. Evaluating protected attributes in isolation risks overlooking how their interactions compound disadvantage in algorithmic systems.

Going forward, future research and system design must address four core challenges: (1) the domain-specific nature of bias, (2) the interactive effects of protected attributes, (3) the variability and robustness of fairness metrics across model configurations, and (4) the real-world consequences of both under- and over-prediction. A fairness framework that is not only statistically rigorous but also socially and ethically responsive will be essential for ensuring equity in high-stakes applications.

Table 7.6: Aggregate mean, standard deviation, and 95% confidence intervals (CI) for fairness metrics and accuracy across all model configurations on the Adult and COMPAS datasets.

Metric	Adult			COMPAS		
	Mean	Std Dev	95% CI	Mean	Std Dev	95% CI
SPD	-0.1376	0.1189	[-0.1837, -0.0915]	0.0529	0.1205	[0.0062, 0.0996]
DI	0.6695	0.3734	[0.5217, 0.8173]	1.1473	0.3325	[1.0157, 1.2789]
EOD	-0.0055	0.0745	[-0.0349, 0.0239]	0.0286	0.1114	[-0.0154, 0.0726]
PED	-0.0503	0.0803	[-0.0821, -0.0185]	0.0426	0.1041	[0.0015, 0.0837]
Acc	0.7926	0.0225	[0.7838, 0.8014]	0.6241	0.0524	[0.6033, 0.6449]

Table 7.7: Subgroup-level mean, standard deviation, and 95% confidence intervals (CI) for fairness metrics on the Adult and COMPAS datasets. Metrics are disaggregated across intersectional group IDs (1–3)

Metric	Group	Adult			COMPAS		
		Mean	Std Dev	95% CI	Mean	Std Dev	95% CI
SPD	1	-0.1789	0.0694	[-0.2063, -0.1515]	0.0234	0.1336	[-0.0314, 0.0782]
	2	-0.0356	0.1066	[-0.0778, 0.0066]	0.1088	0.1234	[0.0580, 0.1596]
	3	-0.1506	0.1143	[-0.1959, -0.1053]	0.0266	0.1976	[-0.0539, 0.1071]
DI	1	0.5599	0.1663	[0.4941, 0.6257]	1.1262	0.4573	[0.9452, 1.3072]
	2	0.9148	0.2584	[0.8026, 1.0270]	1.2757	0.4565	[1.0952, 1.4562]
	3	0.5889	0.2466	[0.4915, 0.6863]	1.0400	0.6039	[0.8013, 1.2787]
EOD	1	-0.0094	0.0656	[-0.0359, 0.0171]	0.0280	0.1531	[-0.0355, 0.0915]
	2	0.0684	0.1012	[0.0259, 0.1109]	0.0727	0.1288	[0.0178, 0.1276]
	3	0.0365	0.1238	[-0.0134, 0.0864]	0.0672	0.1721	[-0.0069, 0.1413]
PED	1	-0.0775	0.0533	[-0.0984, -0.0566]	0.0389	0.1200	[-0.0099, 0.0877]
	2	0.0448	0.1009	[0.0025, 0.0871]	0.0632	0.1117	[0.0166, 0.1098]
	3	-0.0126	0.1095	[-0.0559, 0.0307]	0.0259	0.1550	[-0.0360, 0.0878]
SAcc	1	0.8378	0.0532	[0.8169, 0.8587]	0.6133	0.0965	[0.5741, 0.6525]
	2	0.7626	0.0694	[0.7335, 0.7917]	0.6278	0.0723	[0.5985, 0.6571]
	3	0.7889	0.0981	[0.7501, 0.8277]	0.6022	0.1096	[0.5568, 0.6476]

7.8.1 Multi-Objective and Bi-Objective Optimization: Adult Dataset

To evaluate how different fairness-enhancing strategies balance predictive accuracy with multiple fairness objectives, both multi-objective and bi-objective optimization were applied to the Adult dataset. The optimization considered aggregate-level metrics, including Statistical Parity Difference (SPD), Disparate Impact (DI), Equal Opportunity Difference (EOD), Predictive Equality Difference (PED), and accuracy.

From this analysis, three configurations emerged as non-dominated solutions: `Re+EGR+EO`, `AL+EO`, and `Re+AL+ROC`. Each of these models demonstrated distinct strengths depending on the fairness criteria prioritized.

The `AL+EO` configuration achieved the highest predictive accuracy (0.81) among the three, while still yielding noticeable improvements in group-level fairness. In contrast, the `Re+EGR+EO` model delivered a more comprehensive balance, minimizing both EOD and PED, achieving near-zero values for each, while also reducing SPD and maintaining a moderate Disparate Impact (DI = 0.7236). This model is particularly well-suited for contexts that require balanced trade-offs across multiple fairness dimensions.

Meanwhile, the `Re+AL+ROC` configuration achieved the best Disparate Impact score (DI = 0.7746), making it a strong candidate in scenarios where demographic parity or statistical parity is prioritized, even if it comes at the expense of other fairness metrics.

Under the bi-objective framework, which evaluates trade-offs between accuracy and one fairness metric at a time, the contrasts between configurations became more pronounced. For instance, when jointly optimizing for accuracy and SPD, both `AL+EO` and `Re+EGR+EO` emerged as favorable. `AL+EO` prioritized higher accuracy, while `Re+EGR+EO` provided stronger mitigation of disparity. For EOD, `Re+EGR+EO` stood out by driving disparity nearly to zero without a significant loss in accuracy. When balancing DI with accuracy, `Re+AL+ROC` achieved the best parity outcome, although it showed moderate compromises on other fairness indicators.

Subgroup Fairness Validation (Table 7.3, 7.4, 7.5 Analysis)

The subgroup-level performance reported in Table 7.3, 7.4, 7.5 reinforces the findings from the broader optimization analysis. Notably, the `Re+EGR+EO` (Table 7.5) configuration sustained high levels of accuracy across all demographic intersections, including 0.79 for Female Black individuals and 0.74 for Male Black individuals, two groups that have historically faced systemic disadvantages in economic decision-making scenarios.

The `AL+EO` (Table 7.4) model similarly demonstrated strong subgroup performance, achieving a high accuracy of 0.86 for Female White individuals and 0.73 for Male Black individuals. However, its accuracy for Female Black individuals was slightly lower at 0.72, underscoring a modest trade-off in intersectional fairness compared to `Re+EGR+EO`.

These results affirm that both models are capable of delivering equitable outcomes across multiple protected attribute intersections without causing subgroup performance collapse. In contrast, other configurations such as `DIR+EO` (Table 7.4), while showing favorable results on certain aggregate fairness metrics, exhibited less consistency in subgroup outcomes. This inconsistency highlights a key limitation of single-metric optimization approaches, particularly when applied to intersectional fairness contexts where multiple axes of marginalization interact.

7.8.2 Multi-Objective and Bi-Objective Optimization: COMPAS Dataset

For the COMPAS dataset, the optimization process revealed notable trade-offs between fairness and accuracy, largely due to the domain's sensitivity to unequal error distributions, such as false positives and false negatives. Three configurations were identified as Pareto-optimal when evaluated using aggregate fairness metrics: `EO`, `Re+EGR+EO`, and `DIR+EGR+EO`.

The `EO` model achieved outstanding fairness outcomes, recording a near-zero Equal Opportunity Difference (EOD = 0.0036) and Predictive Equality Differ-

ence (PED = -0.0003), while still preserving an accuracy of 0.65. Meanwhile, the `Re+EGR+EO` model offered comparable fairness performance, albeit with a slightly reduced accuracy of 0.61. However, it showed improvements in Statistical Parity Difference (SPD) and Disparate Impact (DI), making it more balanced across multiple fairness indicators.

Although `DIR+EGR+EO` produced favorable fairness scores, DI = 1.0221 and SPD = 0.0059, its overall accuracy fell to 0.55, which falls below acceptable thresholds for practical use in sensitive domains like criminal justice.

In the context of bi-objective optimization, where accuracy is evaluated alongside each fairness metric individually, the `EO` configuration consistently offered the most effective trade-offs. It was particularly strong in reducing EOD and PED, which are essential for ensuring fairness in error rates among demographic groups in recidivism predictions.

Subgroup Fairness Validation (Table 7.3 Analysis)

The subgroup-level analysis presented in Table 7.3 affirms the effectiveness of the `EO` (Table 7.3) configuration in maintaining equitable performance across demographic groups within the COMPAS dataset. It sustained subgroup accuracies above 0.62 for all groups, achieving its highest accuracy of 0.65 for Female African-American individuals. By contrast, the `Re+EGR+EO` (Table 7.5) model produced a slightly lower accuracy of 0.62 for the same subgroup, though still within an acceptable performance range.

In stark contrast, some configurations that performed well on aggregate metrics, such as `DIR+AL+ROC`, were found to significantly underperform for specific subgroups. For instance, this model reduced the subgroup accuracy for Female African-Americans to just 0.33. Likewise, `DIR+EGR+EO` resulted in subgroup accuracies of 0.55 or lower, further illustrating how aggregate-level improvements can obscure serious harms at the intersectional level.

These results highlight the necessity of evaluating model performance through a

disaggregated lens. Without such analysis, critical disparities may remain hidden, allowing models to perpetuate or worsen unfair treatment of already marginalized populations.

7.8.3 Cross-Dataset Multi-Objective Optimization (MOO)

To assess the generalizability of fairness-enhancing configurations across domains, a multi-objective optimization (MOO) was carried out using both the Adult and COMPAS datasets. The optimization simultaneously considered five conflicting objectives: minimizing Statistical Parity Difference (SPD), Equal Opportunity Difference (EOD), and Predictive Equality Difference (PED); while maximizing Disparate Impact (DI, ideally approaching 1.0) and overall predictive accuracy.

Out of 27 evaluated configurations, three models, `Re+EGR+EO`, `AL+EO`, and `EO`, emerged as Pareto-optimal across both datasets. Among them, `Re+EGR+EO` achieved the most balanced outcomes, maintaining moderate accuracy while reducing disparities across all fairness metrics. `AL+EO` attained the highest mean accuracy but exhibited greater variability in fairness, particularly within the COMPAS dataset. Meanwhile, the `EO` model demonstrated near-ideal EOD and PED values in COMPAS and maintained solid performance in Adult, making it particularly valuable for applications where fairness in error rates is critical.

Other configurations, such as `DIR+EO`, showed strong results within single datasets (e.g., Adult), but underperformed in cross-domain evaluations due to lower subgroup accuracy and fairness stability in COMPAS. These findings emphasize the need to assess fairness strategies not only in isolation but also across diverse real-world contexts.

7.8.4 Cross-Dataset Bi-Objective Optimization

In addition to multi-objective evaluation, bi-objective optimization was performed to explore the trade-offs between accuracy and individual fairness metrics. For each fairness dimension (SPD, DI, EOD, PED), model configurations that most effectively

balanced fairness and predictive performance were identified.

When jointly optimizing accuracy and SPD across both datasets, both `AL+EO` and `Re+EGR+EO` were prominent. `AL+EO` achieved the highest accuracy overall, while `Re+EGR+EO` offered more robust parity improvements. In terms of Disparate Impact (DI), `Re+AL+ROC` performed best on the Adult dataset, though `Re+EGR+EO` showed more consistent cross-domain results. For Equal Opportunity Difference, `EO` delivered strong performance, particularly in COMPAS where it achieved near-zero EOD, while still maintaining fairness in Adult. A similar pattern emerged for PED, with both `EO` and `Re+EGR+EO` achieving strong predictive parity.

Additionally, worst-case subgroup performance was evaluated to assess ethical robustness. The `Re+EGR+EO` model consistently reduced subgroup-level disparities while keeping accuracy above 0.74 for underrepresented groups in Adult and above 0.57 in COMPAS. Although `AL+EO` showed higher average accuracy in Adult, its subgroup performance in COMPAS, especially for Female African-American individuals, was slightly lower. By contrast, the `EO` model maintained stronger minimum subgroup accuracy and consistent treatment in COMPAS, underscoring its reliability as a fairness-oriented baseline across domains.

Together, these results support the value of pipeline-integrated configurations like `Re+EGR+EO`, `EO`, and `AL+EO` as broadly applicable strategies for fairness-aware machine learning. They also stress the importance of evaluating models not only through aggregate performance but by their ability to mitigate harm for the most disadvantaged subgroups across different social domains.

7.9 Health insurance Intersectionality Analysis

This study employs a comprehensive three-stage fairness-aware machine learning pipeline consisting of: (1) a pre-processing method, Disparate Impact Remover (DIR), (2) an in-processing approach, Exponentiated Gradient Reduction (EGR) with XGBoost, and (3) a post-processing technique, Equalized Odds (EO). Evaluation is conducted at the level of intersectional subgroups to assess both performance

and fairness impacts.

7.9.1 Data Preparation and Partitioning

Initial preprocessing involved filtering out instances lacking target labels or key demographic features. To mitigate proxy bias, features highly correlated with protected attributes were excluded. Categorical features were one-hot encoded, and missing values were imputed using the mean for numerical variables and the mode for categorical ones.

To support intersectional fairness analysis, we created a composite attribute (`RACE_GENDER_AGE`) that combines race, gender, and age groupings. This produced ten distinct subgroups. Following domain-specific insights and observed disparities, *White Male Age 65+* was designated as the privileged group, with all other intersections classified as unprivileged.

The dataset was divided into three parts: 50% was allocated for training the base XGBoost classifier and fairness-enhancing models (DIR, EGR, and their combination), 40% was used to evaluate these models and train the EO post-processing mechanism, and the final 10% was held out for evaluating the performance and fairness of the EO-adjusted models.

7.9.2 Fairness Approaches and Evaluation Metrics

This study follows the methodology outlined in Section 4, applying fairness interventions at three stages: pre-processing using Disparate Impact Remover (DIR) (Feldman et al. 2015), in-processing via Exponentiated Gradient Reduction (EGR) (Agarwal et al. 2018), excluding protected attributes from model inputs, and post-processing with Equalized Odds (EO) (Hardt, Price, and Srebro 2016). EGR is paired with an XGBoost classifier, and classification thresholds for EO were tuned per demographic group using a grid search (0.10 to 0.89) to maximize F1 score.

Model performance was evaluated using accuracy, balanced accuracy, F1 score,

recall, and AUC-ROC, reported per intersectional subgroup. Fairness was assessed using four group metrics: Disparate Impact (DI), Statistical Parity Difference (SPD), Equal Opportunity Difference (EOD), and Average Odds Difference (AOD). All fairness metrics were computed using the `AIF360` library (Bellamy et al. 2019) for consistency and reproducibility.

7.9.3 Fairness Results and Analysis

We assessed fairness across ten intersectional subgroups, formed by combining race, gender, and age, using four key metrics: Disparate Impact (DI), Statistical Parity Difference (SPD), Equal Opportunity Difference (EOD), and Average Odds Difference (AOD) (see Table 7.8).

The baseline model, lacking any fairness interventions, showed substantial disparities. For instance, the *Asian Male Age 65+* subgroup recorded DI = 0.33, SPD = -0.57, EOD = -0.55, and AOD = -0.54, signaling pronounced bias and the need for mitigation.

Using the Disparate Impact Remover (DIR) in pre-processing led to modest DI improvements in certain groups (e.g., DI = 0.40 for *Asian Male Age 65+*), but it left other metrics (SPD, EOD, AOD) largely unchanged. This indicates that correcting input distributions alone does not address deeper issues in decision boundary fairness.

Exponentiated Gradient Reduction (EGR), an in-processing method, resulted in more balanced improvements. Across all subgroups, DI values ranged from 0.99 to 1.17, while disparity metrics stayed within -0.01 to 0.15. Notably, DI for *Asian Male Age 65+* rose from 0.34 (baseline) to 0.99, and *Latino Female Age 65+* reached DI = 1.02 and SPD = 0.01. Importantly, these fairness gains were achieved with just a 7% reduction in the base F1 score.

The Equalized Odds (EO) post-processing method, when used alone, was less effective. While it yielded minor fairness gains, especially for EOD and AOD, subgroups like *Asian Male Age 65+* still showed DI = 0.34, suggesting post-hoc adjust-

ments lack impact without prior-stage correction.

Combining DIR and EGR produced a stronger fairness-performance balance. Most subgroups achieved DI between 0.79 and 0.98, SPD between -0.01 and -0.16, and an average F1 score at 94% of the baseline, highlighting the benefit of integrated pre- and in-processing interventions.

The EGR+EO pipeline consistently improved fairness while preserving performance. For example, *Latino Male Age 65+* achieved DI = 0.99, SPD = 0.0, EOD = 0.01, and AOD = 0.02, demonstrating strong fairness alignment with minimal predictive loss.

By contrast, DIR+EO (omitting in-processing) yielded inconsistent results. Not only did fairness metrics stagnate, but performance also declined, such as the F1 score dropping from 0.26 to 0.16 for *Black Male Age 65+*, emphasizing the limits of post-processing when used in isolation.

The full pipeline (DIR + EGR + EO) showed mixed outcomes. Although some metrics improved, the DI for *Asian Male Age 65+* remained suboptimal at 0.78, and subgroup F1 scores dropped in cases like *Black Male Age 65+*. This underscores the risk of overcorrection and the importance of carefully tuning multi-stage mitigation strategies.

Overall, our results show that fairness interventions can mitigate disparities but may compromise performance, particularly if not tailored properly. For instance, EGR+EO improved AOD by an average of 32% across subgroups with 93% F1 score retention, while DIR+EGR maintained 94% of the base F1 and achieved near-parity DI. Conversely, configurations like DIR+EO or DIR+EGR+EO led to either fairness drift or larger drops in accuracy.

These insights affirm that no single strategy is universally optimal. Fairness outcomes are highly dependent on subgroup characteristics, and successful mitigation demands context-specific combinations that are both targeted and adaptive.

Table 7.8: Evaluation of performance and fairness metrics across intersectional subgroups for each pipeline configuration. Negative metric values are indicated in parentheses. BT denotes the Best Threshold used for classification.

Model	Description	BT	Acc	F1-Score	Recall	AUC	DI	SPD	EOD	AOD
Base	White, Female, Age 65+	0.56	0.72	0.26	0.46	0.66	0.99	(0.01)	(0.01)	(0.01)
	Black, Male, Age 65+	0.57	0.74	0.26	0.44	0.67	0.98	(0.02)	(0.03)	-
	Latino, Male, Age 65+	0.59	0.76	0.28	0.42	0.67	0.85	(0.12)	(0.12)	(0.11)
	White, Male, Age under 65	0.56	0.73	0.27	0.48	0.67	0.78	(0.18)	(0.18)	(0.17)
	Latino, Female, Age 65+	0.60	0.79	0.28	0.38	0.68	0.80	(0.17)	(0.16)	(0.17)
	White, Female, Age under 65	0.56	0.73	0.27	0.48	0.67	0.72	(0.23)	(0.23)	(0.24)
	Black, Female, Age 65+	0.58	0.75	0.26	0.41	0.66	0.78	(0.19)	(0.18)	(0.20)
	Asian, Male, Age 65+	0.58	0.75	0.27	0.46	0.66	0.33	(0.57)	(0.55)	(0.54)
	Other	0.57	0.74	0.27	0.47	0.68	0.44	(0.48)	(0.47)	(0.47)
DIR	White, Female, Age 65+	0.55	0.73	0.25	0.43	0.66	1.00	-	-	(0.01)
	Black, Male, Age 65+	0.54	0.72	0.27	0.48	0.67	0.96	(0.03)	(0.04)	(0.02)
	Latino, Male, Age 65+	0.52	0.68	0.27	0.53	0.66	0.83	(0.14)	(0.14)	(0.13)
	White, Male, Age under 65	0.56	0.74	0.27	0.44	0.67	0.73	(0.23)	(0.22)	(0.22)
	Latino, Female, Age 65+	0.56	0.76	0.26	0.42	0.66	0.74	(0.22)	(0.21)	(0.22)
	White, Female, Age under 65	0.58	0.77	0.28	0.39	0.67	0.66	(0.29)	(0.28)	(0.30)
	Black, Female, Age 65+	0.58	0.78	0.27	0.39	0.67	0.70	(0.25)	(0.24)	(0.27)
	Asian, Male, Age 65+	0.59	0.80	0.27	0.35	0.67	0.40	(0.51)	(0.49)	(0.48)
	Other	0.58	0.78	0.27	0.38	0.66	0.43	(0.48)	(0.47)	(0.48)
EGR	White, Female, Age 65+	0.59	0.65	0.25	0.53	0.63	1.01	0.01	0.01	-
	Black, Male, Age 65+	0.67	0.66	0.24	0.50	0.63	1.07	0.05	0.05	0.06
	Latino, Male, Age 65+	0.83	0.72	0.25	0.44	0.63	1.02	0.02	0.01	0.04
	White, Male, Age under 65	0.72	0.67	0.25	0.50	0.63	1.17	0.11	0.11	0.15
	Latino, Female, Age 65+	0.38	0.60	0.24	0.61	0.63	1.02	0.01	0.02	0.02
	White, Female, Age under 65	0.38	0.61	0.28	0.62	0.64	1.15	0.10	0.10	0.12
	Black, Female, Age 65+	0.87	0.79	0.26	0.34	0.65	1.00	-	-	0.01
	Asian, Male, Age 65+	0.63	0.66	0.25	0.49	0.61	0.99	(0.01)	-	0.04
	Other	0.83	0.71	0.24	0.44	0.63	1.05	0.03	0.04	0.05
EO	White, Female, Age 65+	-	0.83	0.20	0.24	-	1.00	-	(0.01)	-
	Black, Male, Age 65+	-	0.82	0.20	0.24	-	0.99	-	(0.01)	0.01
	Latino, Male, Age 65+	-	0.71	0.22	0.34	-	0.81	(0.16)	(0.16)	(0.14)
	White, Male, Age under 65	-	0.71	0.26	0.41	-	0.76	(0.20)	(0.19)	(0.21)
	Latino, Female, Age 65+	-	0.69	0.25	0.37	-	0.77	(0.19)	(0.18)	(0.18)
	White, Female, Age under 65	-	0.68	0.27	0.45	-	0.71	(0.24)	(0.23)	(0.25)
	Black, Female, Age 65+	-	0.71	0.27	0.44	-	0.76	(0.20)	(0.20)	(0.20)
	Asian, Male, Age 65+	-	0.51	0.32	0.66	-	0.34	(0.55)	(0.53)	(0.53)
	Other	-	0.55	0.28	0.59	-	0.44	(0.47)	(0.46)	(0.46)

DIR+EGR	White, Female, Age 65+	0.81	0.71	0.24	0.44	0.63	0.98	(0.01)	(0.01)	(0.02)
	Black, Male, Age 65+	0.35	0.67	0.24	0.50	0.62	0.96	(0.03)	(0.03)	(0.01)
	Latino, Male, Age 65+	0.31	0.65	0.24	0.50	0.62	0.95	(0.03)	(0.03)	(0.01)
	White, Male, Age under 65	0.63	0.68	0.26	0.53	0.64	0.87	(0.09)	(0.09)	(0.07)
	Latino, Female, Age 65+	0.33	0.67	0.27	0.56	0.66	0.92	(0.06)	(0.05)	(0.05)
	White, Female, Age under 65	0.85	0.76	0.27	0.39	0.64	0.85	(0.11)	(0.10)	(0.10)
	Black, Female, Age 65+	0.81	0.71	0.26	0.46	0.64	0.84	(0.12)	(0.11)	(0.11)
	Asian, Male, Age 65+	0.31	0.66	0.25	0.52	0.62	0.79	(0.16)	(0.15)	(0.12)
	Other	0.64	0.68	0.25	0.49	0.62	0.81	(0.14)	(0.13)	(0.13)
EGR+EO	White, Female, Age 65+	-	0.69	0.23	0.51	-	1.01	0.01	0.01	(0.01)
	Black, Male, Age 65+	-	0.69	0.19	0.38	-	1.05	0.03	0.03	0.06
	Latino, Male, Age 65+	-	0.66	0.24	0.44	-	0.99	-	(0.01)	0.02
	White, Male, Age under 65	-	0.73	0.24	0.32	-	1.15	0.10	0.10	0.13
	Latino, Female, Age 65+	-	0.67	0.30	0.47	-	1.00	-	0.01	0.01
	White, Female, Age under 65	-	0.75	0.23	0.38	-	1.14	0.10	0.10	0.10
	Black, Female, Age 65+	-	0.66	0.28	0.49	-	0.98	(0.01)	-	-
	Asian, Male, Age 65+	-	0.62	0.28	0.38	-	0.99	-	(0.01)	0.05
	Other	-	0.68	0.31	0.46	-	1.03	0.02	0.03	0.03
DIR+EO	White, Female, Age 65+	-	0.84	0.20	0.21	-	1.00	-	-	(0.02)
	Black, Male, Age 65+	-	0.81	0.16	0.20	-	0.95	(0.04)	(0.05)	(0.02)
	Latino, Male, Age 65+	-	0.74	0.21	0.29	-	0.81	(0.16)	(0.16)	(0.15)
	White, Male, Age under 65	-	0.71	0.23	0.33	-	0.76	(0.20)	(0.20)	(0.19)
	Latino, Female, Age 65+	-	0.72	0.29	0.39	-	0.78	(0.19)	(0.17)	(0.20)
	White, Female, Age under 65	-	0.68	0.22	0.40	-	0.71	(0.24)	(0.24)	(0.27)
	Black, Female, Age 65+	-	0.72	0.21	0.34	-	0.76	(0.21)	(0.20)	(0.22)
	Asian, Male, Age 65+	-	0.54	0.34	0.55	-	0.40	(0.51)	(0.50)	(0.49)
	Other	-	0.56	0.27	0.52	-	0.46	(0.47)	(0.46)	(0.46)
DIR+EGR+EO	White, Female, Age 65+	-	0.73	0.23	0.46	-	1.00	-	-	(0.02)
	Black, Male, Age 65+	-	0.69	0.18	0.36	-	0.96	(0.03)	(0.03)	0.01
	Latino, Male, Age 65+	-	0.68	0.26	0.43	-	0.93	(0.05)	(0.04)	(0.03)
	White, Male, Age under 65	-	0.66	0.26	0.48	-	0.88	(0.09)	(0.08)	(0.08)
	Latino, Female, Age 65+	-	0.68	0.27	0.44	-	0.93	(0.05)	(0.04)	(0.04)
	White, Female, Age under 65	-	0.67	0.25	0.54	-	0.89	(0.08)	(0.07)	(0.09)
	Black, Female, Age 65+	-	0.66	0.27	0.53	-	0.86	(0.11)	(0.10)	(0.11)
	Asian, Male, Age 65+	-	0.60	0.35	0.55	-	0.78	(0.16)	(0.15)	(0.14)
	Other	-	0.62	0.30	0.52	-	0.82	(0.13)	(0.12)	(0.12)

7.9.4 Key Insights from the health insurance Data Analysis

The empirical findings yield several important observations regarding the relationship between fairness interventions and predictive performance across intersectional subgroups (see Table 7.2).

Insight 1: Multi-stage interventions are more effective than isolated approaches. The results demonstrate that applying fairness strategies at multiple stages of the machine learning pipeline produces more robust improvements in equity. Pipeline combinations such as *DIR + EGR* and *EGR + EO* consistently outperformed single-stage methods across key metrics like DI, SPD, and AOD. These configurations not only mitigated disparities more effectively but also preserved model performance, reinforcing the importance of integrated, pipeline-wide fairness design rather than isolated fixes.

Insight 2: Fairness-performance trade-offs are context-dependent. The impact of fairness interventions on model performance varied significantly across configurations and subgroups. While some combinations, such as *EGR + EO*, delivered strong fairness improvements with minimal reductions in recall or F1 scores, others (e.g., *DIR + EGR + EO*) introduced greater volatility. In particular, performance drops were observed for subgroups like *White Male under 65*, highlighting that trade-offs between fairness and utility are not uniform and must be evaluated on a per-subgroup basis.

Insight 3: Apparent fairness gains can conceal underlying instability. Although DI values for many subgroups converged near the ideal of 1.0 under well-configured pipelines, some results suggested overcompensation. For example, *Latino Male Age 65+* attained a DI of 0.93 under *DIR + EO*, and 0.82 under *DIR + EGR + EO*, which may indicate disproportionate adjustments that do not necessarily reflect fairer outcomes. These findings call attention to the need for comprehensive subgroup audits, as average fairness improvements can obscure model instability or reverse disparities.

Insight 4: Intersectional analysis is essential for identifying compounded

disadvantage. Certain groups, especially *Asian Male Age 65+*, consistently registered low fairness and accuracy scores, even under fairness-aware models. For instance, this subgroup recorded DI values below 0.5 across several configurations, underscoring how intersecting identities can intensify disparities. Single-axis evaluations (e.g., race-only or gender-only) would likely miss such patterns, affirming that intersectional audits are vital for surfacing deeper structural harms.

Insight 5: Post-processing is most effective when built on upstream fairness. Equalized Odds (EO) post-processing yielded the best fairness improvements, such as in EOD and AOD, when used alongside earlier interventions like EGR. When applied alone, EO had limited impact and failed to fully address entrenched bias from earlier stages in the pipeline. This suggests that EO should be treated as a complementary step rather than a standalone fix, most effective when following appropriate pre- or in-processing adjustments.

Taken together, these insights underscore the importance of fairness-aware model development that is both technically comprehensive and socially attentive. Fairness interventions should be tested in combination, tailored to the dataset, and validated across diverse intersectional subgroups to ensure equitable outcomes across the full spectrum of model behavior.

7.10 Guiding Principles for Intersectional Fairness

This study offers both empirical and methodological insight into the challenges of building fair AI systems. Drawing on cross-domain, intersectional, and multi-method evaluations, we present five guiding principles to support the ethical design, implementation, and auditing of fairness-aware machine learning. These principles synthesize empirical evidence with normative reasoning to advance responsible algorithmic practice.

1. **Intersectionality is Essential, Not Optional**

Fairness assessments that ignore the intersection of protected attributes like race and gender risk concealing serious harms to multiply marginalized groups. Our results show that aggregate metrics often obscure subgroup disparities. For instance, in the COMPAS dataset, the overall SPD of 0.2054 (`NONE`) concealed a much higher SPD of 0.2999 for Male African-Americans (see Table 7.2 and Table 7.3). Under the `DIR+AL+ROC` model, accuracy for Female African-Americans dropped to 0.33. Similarly, in the Adult dataset, Female Black individuals had a DI of 0.1409, far below the aggregate DI of 0.3206.

These disparities illustrate the phenomenon of *fairness gerrymandering*, where single-axis metrics give the illusion of fairness while masking harm at intersections. Although current regulations (e.g., the EU AI Act) do not mandate intersectional audits, institutionalizing such evaluations is crucial. Intersectional fairness must be embedded into both auditing tools and model reporting standards.

2. **Fairness–Accuracy Trade-offs are Real but Manageable**

   Tensions between fairness and predictive performance are not only inevitable, but also dependent on application context. In COMPAS, the `EO` model achieved near-perfect EOD (0.0036), but reduced accuracy from 0.68 to 0.65 (Table 7.2). In Adult, the `Re+EGR+EO` strategy improved SPD by 61% with only a minor accuracy loss (0.81 to 0.78).

   Notably, some fairness metrics are mutually incompatible. Improving SPD may degrade EOD or PED, and vice versa. This underscores the importance of aligning fairness goals with domain needs: in criminal justice, EOD may be prioritized due to error sensitivity, while hiring may emphasize statistical parity for representation.

3. **Fairness Interventions Must be Context-Aware**

   Mitigation strategies must align with the type of bias present in the data. The Adult dataset displayed representational disparities, e.g., a DI of 0.1409 for

Female Black individuals, which were effectively addressed through a combination of Reweighing and Exponentiated Gradient Reduction (see Table 7.3).

In contrast, the COMPAS dataset exhibited error-based disparities, such as a high EOD of 0.2906 for Male African-Americans, which responded better to post-processing methods like `EO`. The findings reinforce the value of context-specific strategies: pre-processing handles data imbalance, in-processing imposes fairness during learning, and post-processing corrects output bias. Crucially, subgroup-level evaluation is essential to avoid reinforcing structural inequities.

4. **Subgroup Auditing Must Be Standard Practice**

   Relying only on aggregate metrics can mask subgroup harms. In COMPAS, the `Re+AL+ROC` model achieved an aggregate DI near 1.0, yet Female African-Americans experienced a DI of 1.3582 (Table 7.2 and Table 7.3). Similarly, Female Black individuals in the Adult dataset often had the lowest DI across models.

   This highlights the need for disaggregated analysis. Fairness assessments should report subgroup-specific metrics alongside global scores. Moreover, systematically identifying and ranking the most disadvantaged subgroups ensures that fairness interventions do not ignore those most affected.

5. **Fairness Should Prioritize the Most Disadvantaged**

   Achieving fairness is not just about balancing outcomes between groups. It also requires an ethical commitment to improving outcomes for those historically most harmed. This aligns with Binns' theory of prioritizing the worst-off in algorithmic decision-making(Binns 2018).

   Our results showed that average fairness improvements often failed to uplift the most marginalized, such as Black women, who continued to suffer low accuracy and poor SPD in COMPAS. Models should therefore embed subgroup disadvantage into their design and evaluation pipelines. In real-world use,

continuous monitoring of subgroup outcomes is critical to detecting drift and re-emerging disparities.

These principles emphasize the importance of fairness frameworks that are not only technically sound, but also ethically robust and contextually grounded. To advance equity in AI, we must move beyond isolated metrics and focus on how systems affect the most marginalized, measuring fairness not only by parity, but by justice.

7.11 Summary

This chapter examined intersectional subgroup disparities across multiple datasets, Adult Income, COMPAS, and health insurance, using a range of fairness-enhancing techniques applied at different stages of the fairness pipeline. Across domains, the analysis consistently revealed that aggregate fairness metrics often obscure deeper harms affecting multiply marginalized groups. Subgroups such as Black women and older Asian men experienced persistent disparities in both predictive performance and fairness metrics, even when overall model outcomes appeared balanced.

Multi-stage mitigation strategies, particularly combinations like `Re+EGR+EO` and `DIR+EGR`, demonstrated stronger and more consistent improvements across fairness dimensions without substantial losses in predictive accuracy. However, no single intervention proved universally optimal; effectiveness varied depending on dataset characteristics, bias types, and the intersectional composition of subgroups.

These findings reinforce the critical importance of disaggregated evaluation and pipeline-level fairness interventions. They also highlight the need to prioritize the most disadvantaged groups through targeted mitigation strategies. This sets the stage for the next chapter, which focuses on applied use cases where intersectional fairness principles are operationalized in real-world deployment settings.

Publication(s) Arising from this Chapter

The work presented in this chapter has been published (or submitted) in the following outlets:

1. Michael Farayola et al. (2025). "Beyond Aggregate Fairness: Intersectional Auditing Across the AI Fairness Pipeline". In: *AI and Ethics*

2. Michael Mayowa Farayola, Shane Kennedy, et al. (2025). "Intersectional Fairness in Healthcare AI: A Pipeline-Wide Evaluation of Multi-Stage Mitigation Strategies". In

These publication(s) reflect the main contributions of this chapter and provide further technical details, extended results, and peer-reviewed validation of the methods and findings.

Chapter 8

Discussion and Conclusion: Contributions, Implications, and Future Directions

8.1 Introduction

This chapter consolidates the insights developed throughout this thesis. It demonstrates how the research carried out in this thesis advances fairness-aware AI systems in recidivism prediction and related high-stakes domains. Earlier chapters served distinct purposes, establishing conceptual foundations (Chapters 2 and 3), outlining the methodological design (Chapter 4), and presenting empirical findings on integrated fairness pipelines (Chapter 5), oversampling strategies (Chapter 6), and intersectional auditing (Chapter 7). In this chapter, we weave all these strands into a unified narrative that directly addresses the guiding research questions and situates the contributions within broader academic and practical contexts.

This chapter begins by revisiting the research questions that framed the study carried out in this thesis. Each question is linked to the relevant chapters (investigations) and the conclusions or insights that emerged from them, ensuring continuity with the aims and hypotheses articulated in Chapter 1. The discussion then high-

lights cross-cutting insights that transcend individual studies, demonstrating how the results collectively strengthen the understanding of fairness in AI within the criminal justice system, particularly in predicting recidivism risk and beyond.

The contributions of the thesis are as follows, and these are organised into two categories: (1) contributions to knowledge, including theoretical frameworks, empirical evidence, and methodological innovations; and (2) contributions to practice, covering implications for policymakers, practitioners, and system developers. A contribution table summarises these advances in a clear and structured manner.

This chapter also examines the ethical and societal implications of the research, as recidivism prediction directly shapes decisions about liberty, rehabilitation, and justice. Similarly, in domains such as finance and healthcare, technical progress must go hand in hand with accountability, legitimacy, and trust. The findings, therefore, position the work within broader debates on trustworthy and justice-aware AI across multiple high-stakes contexts.

Limitations of the research are then acknowledged, including constraints in data, methods, evaluation metrics, and external generalisability. By recognising these boundaries, the chapter provides and sets the stage for future work. Building on these limitations and insights, the discussion outlines directions for further research, focusing on methodological extensions, cross-domain applications, and the integration of governance.

This chapter concludes with a final synthesis that restates the key contributions and reflects on the overarching message of the thesis: achieving fairness in AI requires integrated, intersectional, and socio-technical approaches that combine rigorous empirical methods with ethical responsibility.

8.2 Research Questions Revisit

This section revisits the research questions introduced in Chapter 1 and demonstrates how the research across the thesis addressed them. Each research question is discussed in relation to the empirical and conceptual findings, highlighting how

the results advance the understanding of fairness in artificial intelligence and its application to recidivism prediction and beyond.

RQ1: What are the key ethical and trustworthiness challenges in AI-based recidivism prediction?

Chapters 2 and 3 examined the ethical, technical, and socio-technical challenges of applying AI to recidivism prediction regarding the trustworthiness of AI. The concept of Trustworthy AI is explored, and the analysis further narrows in on fairness as a significant concern for stakeholders in the criminal justice system. Fairness is a multifaceted concept, encompassing competing definitions and often leading to metric incompatibilities. The findings highlighted the fairness-accuracy trade-offs, the inadequacy of relying solely on single fairness notions or fairness-enhancing techniques, and the need to situate fairness within the broader principles of trustworthy AI. In addition, this thesis underscored the ethical imperative of intersectionality, showing that evaluations based solely on single attributes, such as race or gender, conceal the harms faced by individuals at the intersection of multiple marginalized identities. These insights demonstrated that ethical and trustworthiness challenges extend beyond algorithmic performance and must include governance, transparency, and participatory practices.

RQ2: Can integrated fairness interventions across different stages of AI development enhance fairness and predictive accuracy in recidivism models?

Chapter 5 and 7 demonstrated that integrated interventions across pre-processing, in-processing, and post-processing stages consistently improved fairness outcomes compared to single-stage methods. The results showed that integrating fairness-enhancing techniques produced limited bias across multiple fairness metrics while maintaining acceptable predictive accuracy. Through many-objective and bi-objective

optimisation, the study identified Pareto-efficient configurations that made fairness-accuracy trade-offs explicit and transparent. These findings established integrated pipelines as a more robust approach than isolated interventions, offering stakeholders a principled method for selecting models aligned with context-specific priorities.

RQ3: Can fairness-aware oversampling techniques mitigate systemic biases while maintaining predictive accuracy across diverse recidivism datasets?

Chapter 6 provided a systematic evaluation of fairness-aware oversampling strategies. The results indicated that oversampling strategies can reduce bias arising from class imbalance and subgroup underrepresentation; however, its effectiveness depends strongly on dataset characteristics, classifier families, and the type of strategy applied. For example, sensitive-attribute-based oversampling improved fairness in some contexts but risked introducing synthetic bias and overfitting in others. Confidence-interval analyses further revealed that not all observed improvements were statistically robust. These results offered guidance on when oversampling strategies enhance fairness, when they should be applied with caution, and how they interact with different model architectures.

RQ4: To what extent do the proposed approaches generalise beyond criminal justice to other high-stakes domains such as healthcare?

Chapter 7 extended the analysis beyond recidivism by applying the integrated framework while incorporating intersectionality across different domains, including finance and healthcare datasets. The cross-domain experiments demonstrated that fairness pipelines and intersectional auditing remained valuable in this new context, revealing subgroup disparities that aggregate metrics had previously concealed. At the same time, the findings revealed important limitations: fairness interventions that

improved outcomes in one dataset did not always transfer seamlessly to another, and contextual factors shaped the effectiveness of each method. These results demonstrated both the portability and the limits of the proposed approaches, underscoring the need for domain-sensitive governance when transferring fairness methods across high-stakes applications.

In summary, revisiting the research questions help identify the ethical and trustworthiness issues that underpin recidivism prediction, demonstrating the advantages of integrated fairness pipelines, clarifying the role and risks of oversampling, and testing the generalisability of these approaches across domains. Collectively, these findings confirm that achieving fairness in AI requires integrated, data-sensitive, and intersectionally aware interventions embedded within a socio-technical framework of governance and accountability.

8.3 Cross-Chapter Insights

The findings presented in Chapters 5 to 7 reveal a set of cross-cutting insights that extend beyond the individual investigations. Considering the results collectively, we can identify recurring themes, methodological patterns, and conceptual lessons that clarify how fairness-aware AI systems should be designed and evaluated for recidivism prediction and other high-stakes domains.

8.3.1 Integration Outperforms Isolation

The experiments in Chapter 5 demonstrated that fairness interventions produce stronger and more consistent improvements when applied in integration across multiple stages of the pipeline. While single-stage methods provided local benefits, integrated pipelines delivered broader fairness gains with only modest costs to predictive accuracy. This insight suggests that research and practice should move beyond evaluating fairness methods in isolation and focus instead on how complementary techniques can work together. However, it is essential to note that some integrations

may compromise algorithmic fairness, thereby affecting trust.

8.3.2 Data Characteristics Shape Fairness Outcomes

The analysis of the COMPAS, RisCanvi, Adult Income and Irish Insurance datasets revealed that dataset characteristics, such as subgroup size, class imbalance, and distributional irregularities, exert a strong influence on fairness outcomes. In some cases, interventions that improved fairness on one dataset produced negligible or even negative effects on another. These observations highlight the importance of diagnosing dataset properties before applying fairness methods and confirm that no single technique can guarantee equitable results across all contexts.

8.3.3 Intersectional Auditing is Indispensable

Chapter 7 demonstrated that aggregate fairness metrics often mask persistent harms to intersectional subgroups. Even when overall disparities decreased, certain combinations of protected attributes (such as race and gender) continued to experience significant disadvantage. Intersectional auditing therefore emerges as an essential practice, ensuring that fairness evaluations account for those most affected by systemic inequities and preventing misleading conclusions based on averaged results.

8.3.4 Optimisation as a Design Instrument

The optimisation framework introduced in Chapter 5 illustrated how many-objective and bi-objective optimisation can make fairness-accuracy trade-offs explicit. By identifying Pareto-efficient frontiers, stakeholders can select models according to context-specific priorities rather than relying on arbitrary thresholds. Optimisation thus operates not only as a computational tool but also as a design instrument that structures deliberation around competing values.

8.3.5 Governance Anchors Technical Advances

Across all studies, the findings underscored the importance of integrating technical improvements into governance structures. Transparency in reporting, accountability mechanisms, and opportunities for participatory engagement emerged as non-negotiable elements of trustworthy deployment. Although the thesis did not explore every principle of trustworthy AI, it positions the research as a meaningful contribution to enhancing trust in AI systems. Without such socio-technical anchors, technical fairness interventions risk failing to secure legitimacy in real-world applications.

In summary, these insights confirm that fairness in AI cannot be achieved through isolated technical solutions or context-free methodologies. Instead, it requires integrated pipelines, sensitivity to data characteristics, intersectional auditing, optimisation-guided decision-making, and robust governance frameworks. These themes provide a unifying perspective that connects the individual contributions of the thesis into a coherent approach for advancing fairness-aware and trustworthy AI.

8.4 Overview of Contributions

This section highlights the contributions of the thesis, distinguishing between theoretical and empirical advances to the body of knowledge and practical implications for system developers, policymakers, and other stakeholders. Table 8.1 provides a structured summary of these contributions, which are elaborated in the subsections that follow.

Table 8.1: Summary of thesis contributions across research and practice.

Area	**Contribution**
Fairness Integration	Proposed and implemented novel integration strategies combining fairness techniques across pre-, in-, and post-processing stages (M. M. Farayola, Bendechache, Saber, et al. 2024b).
Optimization	Applied multi-objective and bi-objective optimization (e.g., Pareto front) to identify fairness–accuracy trade-offs and optimal model configurations (M. M. Farayola, Bendechache, Saber, et al. 2024a).
Experimental Design	Designed and executed comprehensive benchmarking using COMPAS, RisCanvi, Adult Income and private Irish Insurance datasets with multiple models and fairness metrics(M. M. Farayola, Tal, Saber, et al. 2025; M. Farayola et al. 2025).
Novel Curated Dataset Use	Curated and translated the RisCanvi dataset from Spanish to English, enabling broader use in fairness research (M. M. Farayola, Tal, Saber, et al. 2025).
Trustworthy AI Framework	Proposed an extension of the EU requirements to enhance the ethical and trustworthy use of AI in predicting recidivism risk within the criminal justice system (M. M. Farayola, Tal, Connolly, et al. 2023).
Justice-Aware AI Fairness Framework	Proposed a justice-aware AI fairness framework emphasizing intersectionality, socio-technical context, and ethical design (M. M. Farayola, Malika, et al. 2026).
Survey/Review Work	Mapped fairness techniques to AI pipeline phases; conducted systematic literature reviews on fairness and trustworthiness in recidivism models (M. M. Farayola, Tal, Malika, et al. 2023).
Algorithmic Tooling Enhancement	Redesigned AIF360's DIR to support multi-valued protected attributes (intersectionality) and preserve group identity during feature repair (M. Farayola et al. 2025).
Robust Debiasing Model	Redesigned AIF360's AL to handle intersectionality evaluation and multi-class protected attributes, included softmax adversary output, and validated against data irregularities for stable training (M. Farayola et al. 2025).
Bias-Mitigation Oversampling Approach	Evaluated fairness-aware oversampling strategies addressing class imbalance and subgroup underrepresentation, offering empirical insights into their impact on fairness, accuracy, and model stability across diverse recidivism datasets (M. M. Farayola, Bendechache, Takfarinas, et al. 2025).
Intersectionality Framework	Proposed and evaluated an intersectionality-aware framework integrated with fairness-enhancing techniques across the fairness pipeline for systematic fairness and accuracy evaluation across intersectional subgroups (M. Farayola et al. 2025; M. M. Farayola, S. Kennedy, et al. 2025).

8.4.1 Contribution to the Body of Knowledge

The thesis advances the scholarly literature in several key ways:

1. **Integrated Fairness Pipelines.** The research demonstrates that fairness interventions yield broader and more reliable outcomes when integrated across pre-, in-, and post-processing stages. This finding challenges the dominant practice of testing mitigation techniques in isolation and provides a new perspective on how to combine methods for maximum effect.

2. **Optimization as a Framework for Fairness.** By applying many-objective and bi-objective optimization, the thesis operationalises fairness as a process of navigating trade-offs rather than selecting a single metric. This contribution reframes fairness research as a multi-criteria problem and provides methodological clarity through the explicit mapping of Pareto-efficient solutions.

3. **Empirical Benchmarks and Experimental Design.** The benchmarking experiments on COMPAS and RisCanvi datasets, using a diverse set of classifiers and fairness metrics, expand the empirical base for evaluating fairness-aware machine learning. The translation and curation of the RisCanvi dataset into English further contribute a resource for the community, broadening participation and comparability in fairness research.

4. **Intersectional Framework.** The thesis introduces a justice-aware AI fairness framework that foregrounds intersectionality and socio-technical context. This contribution shifts fairness research away from narrow, single-axis approaches and toward a framework that aligns more closely with principles of justice and equity.

5. **Algorithmic Tooling and Robust Models.** By extending AIF360's DIR and AL modules, the thesis enhances the state of algorithmic tooling available to researchers. These extensions make it possible to handle multi-valued and

multi-class protected attributes and to train more stable adversarial debiasing models in the presence of data irregularities.

6. **Survey and Literature Mapping.** The systematic reviews and mappings conducted in this work clarify how fairness techniques align with the stages of the machine learning pipeline and how recidivism models can be evaluated for trustworthiness. These surveys synthesise fragmented literature and provide a foundation for future studies.

8.4.2 Contribution to Practice

Beyond theoretical and empirical advances, the thesis also delivers practical contributions:

1. **Guidance on Pipeline Design.** The integration strategies and optimisation frameworks provide actionable guidance for practitioners who must balance fairness with accuracy in real-world deployments. The findings help practitioners identify when and how to combine methods for maximum fairness impact.

2. **Evaluation Protocols.** The thesis advocates for the routine use of disaggregated evaluations, confidence intervals, and intersectional auditing as standard practice. These recommendations improve transparency and accountability by preventing misleading conclusions based on aggregate metrics.

3. **Dataset Contribution.** By curating and translating the RisCanvi dataset (RisCanvi GitHub Repository), the thesis provides the community with a valuable resource for testing fairness-aware methods in recidivism prediction. This practical output supports replication, comparability, and broader adoption of fairness evaluations.

4. **Governance Recommendations.** By extending EU requirements for trustworthy AI, the research provides concrete guidance for policymakers and institutions deploying AI in criminal justice. These recommendations foreground

transparency, accountability, and participatory engagement as essential safeguards in high-stakes contexts.

5. **Data-Level Interventions.** The evaluation of fairness-aware resampling strategies, particularly for intersectional subgroups, informs practitioners about the potential and risks of data-level interventions. This insight helps prevent misuse of oversampling techniques while identifying contexts where they can add value.

8.5 Ethical and Societal Implications

The findings of this thesis carry significant ethical and societal implications because recidivism prediction directly influences decisions about liberty, rehabilitation, and justice. Technical progress alone cannot guarantee fair or legitimate outcomes; fairness-aware models must operate within a socio-technical ecosystem that addresses accountability, transparency, and participation. This section reflects on the broader implications of the research and situates the contributions within ongoing debates about trustworthy and justice-aware artificial intelligence.

8.5.1 Fairness Beyond Metrics

The research highlights the danger of treating fairness as a property captured by a single metric. Aggregate performance measures frequently masked harms to smaller or intersectional subgroups, as demonstrated in Chapter 7. This finding underscores an ethical imperative: evaluations must prioritise those most disadvantaged by systemic inequities rather than relying on population averages. A model that appears fair overall can perpetuate serious harms when its performance systematically disadvantages specific communities.

8.5.2 Justice-Aware Approaches

By introducing an intersectional fairness framework, the thesis connects algorithmic fairness to broader notions of social justice. Ethical AI requires not only technical adjustments but also explicit recognition of historical and structural inequalities. Incorporating intersectionality ensures that fairness interventions address the lived realities of those who face overlapping forms of disadvantage. This approach aligns with justice-aware perspectives that call for centring the voices and experiences of the most affected.

8.5.3 Trust, Legitimacy, and Accountability

Trust in recidivism prediction models depends on more than predictive accuracy or technical fairness improvements. Institutions must demonstrate accountability by making models transparent, explainable, and open to scrutiny. The research reinforces the need for governance mechanisms that allow affected individuals to contest outcomes, demand redress, and understand the basis of predictions. Without such accountability structures, even technically advanced models risk undermining institutional legitimacy and exacerbating distrust in the justice system.

8.5.4 Participatory Engagement

The findings emphasise the importance of including impacted stakeholders, such as formerly incarcerated individuals, community organisations, and legal practitioners, in the design, evaluation, and governance of AI systems. Participation ensures that fairness interventions align with community values and address real-world concerns. While this thesis concentrated on technical experimentation, its justice-aware framework points to participatory engagement as a necessary condition for ethically sound deployment.

8.5.5 Societal Risks of Misuse

The research also highlights the risks of misuse or overreliance on fairness interventions. Technical improvements might create a false sense of security if policymakers treat them as substitutes for broader criminal justice reforms. Fairness-aware AI should support, not replace, efforts to address systemic biases in policing, sentencing, and rehabilitation. Ethical deployment therefore requires situating algorithmic interventions within larger reforms aimed at reducing structural inequality.

In summary, the ethical and societal implications of this research extend beyond technical performance. The thesis demonstrates that fairness requires intersectional auditing, justice-aware frameworks, and participatory governance. It further shows that ethical AI must empower affected communities, safeguard accountability, and resist being used as a veneer for systemic inequities. By embedding technical innovations within these broader ethical commitments, the research contributes to building AI systems that are not only more accurate but also more just and trustworthy.

8.6 Limitations

Every research project operates within certain boundaries, and recognising these limitations is essential for transparency and for guiding future work. The contributions of this thesis should therefore be understood in light of several constraints relating to scope, methodology, evaluation, and external validity.

8.6.1 Scope of Data and Tasks

The thesis focused primarily on tabular datasets and binary classification tasks, specifically in the context of recidivism prediction. While this scope allowed for controlled and comparable experimentation, it also restricted the applicability of the findings to other modalities such as text, images, or time-series data. Multi-label and multi-task problems were not explored, which limits the generalisability of the results to more complex prediction settings.

8.6.2 Absence of Real-World Feedback Loops

The analyses treated data as static and did not account for feedback effects that arise when model predictions influence future data collection and institutional behaviour. In real-world deployment, risk assessment tools can shape policing, sentencing, and parole decisions, which in turn affect the distribution of future cases. By not modelling these dynamic feedback loops, the research does not capture long-term systemic impacts of algorithmic interventions.

8.6.3 Fairness Metrics and Statistical Constraints

The evaluation relied on widely recognised fairness metrics, many of which are mathematically incompatible with one another. Trade-offs between equalised odds, demographic parity, and calibration remain unresolved, and the study did not attempt to reconcile these incompatibilities. Moreover, certain subgroup estimates were affected by small sample sizes, which introduced statistical variability and widened confidence intervals. These constraints limited the strength of some intersectional conclusions.

8.6.4 Oversampling within Integrated, Intersectional Pipelines

The thesis evaluated fairness-aware oversampling comprehensively in Chapter 6 and then integrated selected oversampling strategies into the Chapter 7 intersectional experiments. This integration enabled an assessment of compound effects at the subgroup level, where oversampling interacted with fairness-enhancing methods and evaluation procedures within the intersectional auditing framework. The design exposed how data-centric interventions can amplify or attenuate mitigation effects for multiply marginalised subgroups.

Despite this integration, several limitations remain. The experiments did not exhaust all possible combinations of oversampling strategies, integration points, and model families; they focused on representative configurations that balanced methodological clarity with empirical depth. Oversampling also carries risks of

distributional shift and synthetic bias, particularly under severe sparsity for small intersectional groups. While the analysis employed confidence intervals and disaggregated reporting to monitor these risks, further robustness checks (e.g., sensitivity analyses across random seeds, alternative synthetic-sample generators, and stricter regularisation) would strengthen conclusions. Finally, because the integration targeted datasets with specific intersectional sparsity patterns, transferability to other domains and demographic structures warrants additional study.

8.6.5 Generalisability Across Contexts

Although the thesis extended its analysis to the healthcare domain, the majority of experiments were conducted on recidivism datasets. The cultural, legal, and institutional characteristics of criminal justice systems shaped both the data and the interpretation of fairness interventions. As a result, findings may not fully generalise to other sectors or jurisdictions where fairness concerns manifest differently.

8.6.6 Computational and Resource Constraints

The scale of experiments was bounded by available computational resources. Although benchmarking covered multiple models and fairness metrics, the evaluation could not exhaustively test all possible fairness interventions or optimisation strategies. These resource constraints limited the exploration of larger-scale, real-time, or industrial-level applications.

By acknowledging these limitations, the thesis provides a clear boundary for interpreting its results. These constraints do not diminish the contributions but instead highlight opportunities for further exploration. Subsequent research can expand into additional modalities, model dynamic feedback loops, develop methods for robust intersectional inference, and test interventions across diverse contexts and domains.

8.7 Future Directions

The limitations identified in the preceding section, together with the insights gained across Chapters 4 to 7, suggest several promising directions for future research. Advancing fairness-aware and trustworthy artificial intelligence requires extending the methodological, empirical, and socio-technical contributions of this thesis into new domains and applications. The following research avenues build directly on the work presented here.

8.7.1 Expanding Modalities and Tasks

Future work should extend the integrated fairness pipelines beyond tabular data and binary classification. Text, image, and time-series modalities present unique fairness challenges, such as biased word embeddings, skewed visual representations, and irregular temporal sampling. Similarly, multi-label and multi-task learning settings introduce complex dependencies between outputs that may amplify disparities if left unexamined. Evaluating fairness interventions across these modalities and tasks will test the robustness of the framework and expand its applicability to a wider range of high-stakes domains.

8.7.2 Modelling Dynamic Feedback Loops

Recidivism prediction models, like many decision-support tools, do not operate in isolation. Their outputs influence institutional practices such as policing, sentencing, and parole, which in turn shape future datasets. A critical avenue for future research lies in modelling these feedback loops using causal inference, system dynamics, or agent-based simulations. Incorporating these dynamics would make it possible to evaluate not only short-term fairness outcomes but also long-term systemic impacts, including potential reinforcement of structural inequalities.

8.7.3 Robust Intersectional Inference

The results of Chapter 7 highlighted the risks of small sample sizes in intersectional subgroup analyses. Future work should therefore explore statistical methods that provide more reliable inference under data sparsity. Hierarchical Bayesian models, partial pooling, or resampling strategies with uncertainty quantification could strengthen confidence in subgroup-level fairness assessments. Embedding these approaches into fairness evaluation protocols would allow practitioners to report not only point estimates but also uncertainty ranges, enhancing transparency and reliability.

8.7.4 Integrating Data-Centric and Pipeline Interventions

While Chapter 7 integrated oversampling strategies with fairness interventions, further work is required to explore the full design space of data-level and pipeline-level combinations. Research could examine how different oversampling algorithms interact with adversarial debiasing, reweighting, or post-processing adjustments, and whether safeguards such as regularisation or stability constraints can mitigate risks of synthetic bias. More systematic experimentation across datasets and model families would clarify best practices for deploying hybrid approaches.

8.7.5 Cross-Domain and Cross-Jurisdictional Studies

The extension to healthcare in Chapter 7 provided evidence of partial generalisability but also revealed domain-specific limitations. Future research should systematically investigate fairness interventions across multiple domains—such as education, credit scoring, or employment—as well as across jurisdictions with distinct legal and cultural contexts. Such studies would identify the boundary conditions of fairness methods, highlight context-dependent adaptations, and inform guidelines for responsible cross-domain transfer.

8.7.6 Operationalising Participatory and Challenge Mechanisms

The justice-aware framework advanced in this thesis calls for participatory engagement and accountability structures, but the empirical studies did not directly incorporate participatory methods. Future work should operationalise participatory co-design with affected communities, evaluate the effectiveness of challenge mechanisms, and measure their impact on trust and legitimacy. Tools such as model cards, datasheets, and participatory audits could be systematically tested as part of governance interventions to complement technical fairness measures.

8.7.7 Scaling and Resource Considerations

Finally, future work should address scalability and resource efficiency. Industrial-scale deployments often involve large datasets, complex pipelines, and real-time constraints. Research into computationally efficient fairness interventions, approximate optimisation methods, and distributed training approaches would make fairness-aware pipelines more practical for large-scale applications.

8.8 Summary

This chapter brought the thesis to a close by revisiting the research questions, synthesising cross-chapter insights, articulating contributions to knowledge and practice, reflecting on ethical and societal implications, acknowledging limitations, and outlining directions for future work. Together, these sections demonstrate how the thesis advances the study and practice of fairness-aware machine learning in recidivism prediction and related high-stakes domains.

The research questions framed in Chapter 1 guided the investigation and received comprehensive answers. The work identified core ethical and trustworthiness challenges in recidivism prediction, demonstrated the advantages of integrated fairness pipelines, clarified the benefits and risks of fairness-aware oversampling, and tested

the generalisability of the proposed approaches in the healthcare domain. Each question was addressed through empirical experiments, methodological innovation, and critical analysis, confirming the thesis's original aims.

Cross-chapter insights revealed unifying themes: integration outperforms isolation, dataset characteristics strongly shape fairness behaviour, intersectional auditing is indispensable, optimisation can serve as a design instrument for balancing competing goals, and governance anchors technical advances in broader socio-technical commitments. These themes tie together the individual contributions and provide a framework for fairness-aware AI research and practice.

The contributions of the thesis extend both the body of knowledge and professional practice. The research introduced integrated fairness pipelines, optimisation-guided model selection, and intersectional auditing frameworks; curated and translated the RisCanvi dataset; enhanced algorithmic fairness toolkits; and mapped fairness interventions across the machine learning pipeline. In practice, the work offers guidance for pipeline design, protocols for evaluation, governance recommendations, and resources for replication and broader adoption. Collectively, these contributions demonstrate novelty, rigour, and impact across theory, method, and application.

The ethical and societal implications emphasise that fairness in AI cannot be reduced to metrics alone. True progress requires justice-aware frameworks that recognise intersectionality, embed transparency and accountability, and involve participatory engagement. At the same time, the research acknowledges its limitations in scope, methodology, and external validity, providing clear boundaries for interpreting the findings.

Future research directions build directly on these limitations and insights. Promising avenues include extending fairness pipelines to new modalities and tasks, modelling feedback loops, strengthening intersectional inference, combining data-centric and pipeline interventions, conducting cross-domain and cross-jurisdictional studies, and operationalising participatory and governance mechanisms. These directions set the agenda for a continued programme of fairness research that remains empirically

grounded and ethically accountable.

In conclusion, the thesis demonstrates that achieving fairness in AI requires moving beyond isolated technical fixes. It requires integrated, intersectional, and socio-technical approaches that balance accuracy with justice, optimise competing objectives transparently, and embed accountability into governance structures. By combining rigorous empirical investigation with ethical reflection, the research contributes to building AI systems that are not only more accurate but also more equitable, trustworthy, and legitimate.

Bibliography

Chawla, Nitesh V et al. (2002). "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* 16, pp. 321–357.

Batista, Gustavo EAPA, Ana LC Bazzan, Maria Carolina Monard, et al. (2003). "Balancing training data for automated annotation of keywords: a case study." In: *Wob* 3, pp. 10–18.

Batista, Gustavo EAPA, Ronaldo C Prati, and Maria Carolina Monard (2004). "A study of the behavior of several methods for balancing machine learning training data". In: *ACM SIGKDD explorations newsletter* 6.1, pp. 20–29.

Estabrooks, Andrew, Taeho Jo, and Nathalie Japkowicz (2004). "A multiple resampling method for learning from imbalanced data sets". In: *Computational intelligence* 20.1, pp. 18–36.

Han, Hui, Wen-Yuan Wang, and Bing-Huan Mao (2005). "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning". In: *International conference on intelligent computing*. Springer, pp. 878–887.

Dwork, Cynthia (2006). "Differential privacy". In: *International colloquium on automata, languages, and programming*. Springer, pp. 1–12.

Bonta, James and Donald A Andrews (2007). "Risk-need-responsivity model for offender assessment and rehabilitation". In: *Rehabilitation* 6.1, pp. 1–22.

Dwork, Cynthia (2008). "Differential privacy: A survey of results". In: *International conference on theory and applications of models of computation*. Springer, pp. 1–19.

LEARNING, TASET SHIFT IN MACHINE (2009). *Dataset shift in machine learning*.

Cadigan, Timothy P and Christopher T Lowenkamp (2011). "Implementing risk assessment in the federal pretrial services system". In: *Fed. Probation* 75, p. 30.

Dwork, Cynthia et al. (2012). "Fairness through awareness". In: *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226.

Kamiran, Faisal and Toon Calders (2012). "Data preprocessing techniques for classification without discrimination". In: *Knowledge and information systems* 33.1, pp. 1–33. DOI: https://doi.org/10.1007/s10115-011-0463-8.

Kamiran, Faisal, Asim Karim, and Xiangliang Zhang (2012). "Decision theory for discrimination-aware classification". In: *2012 IEEE 12th international conference on data mining*. IEEE, pp. 924–929. DOI: https://doi.org/10.1109/ICDM.2012.45.

Connolly, Regina (2013). "Trust in commercial and personal transactions in the digital age". In: *The Oxford Handbook of Internet Studies*. OUP Oxford, pp. 1–22. DOI: https://doi.org/10.1093/oxfordhb/978019958074.013.0013.

Donnellan, Michael (2013). "Irish Prison Service Recidivism Study". In: *Irish Prison Service, Dublin*.

Monahan, John and Jennifer L Skeem (2013). "Risk redux: The resurgence of risk assessment in criminal sanctioning". In: *Fed. Sent'g Rep.* 26, p. 158. DOI: https://doi-org.dcu.idm.oclc.org/10.1525/fsr.2014.26.3.158.

Szegedy, Christian et al. (2013). "Intriguing properties of neural networks". In: *arXiv preprint arXiv:1312.6199*.

Berk, Richard and Justin Bleich (2014). "Forecasts of violence to inform sentencing decisions". In: *Journal of Quantitative Criminology* 30.1, pp. 79–96. DOI: https://doi.org/10.1007/s10940-013-9195-0.

Goodfellow, Ian J, Jonathon Shlens, and Christian Szegedy (2014). "Explaining and harnessing adversarial examples". In: *arXiv preprint arXiv:1412.6572*.

Edwards, Harrison and Amos Storkey (2015). "Censoring representations with an adversary". In: *arXiv preprint arXiv:1511.05897*. URL: https://iclr.cc/archive/www/doku.php%3Fid=iclr2016:main.html.

Feldman, Michael et al. (2015). "Certifying and removing disparate impact". In: *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 259–268.

Amodei, Dario et al. (2016). "Concrete problems in AI safety". In: *arXiv preprint arXiv:1606.06565*.

Angwin, Julia et al. (2016a). *How We Analyzed the COMPAS Recidivism Algorithm*. URL: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

— (2016b). "Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks". In: *ProPublica*. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Desmarais, Sarah L, Kiersten L Johnson, and Jay P Singh (2016). "Performance of recidivism risk assessment instruments in US correctional settings". In: *Psychological services* 13.3, p. 206. DOI: https://doi.org/10.1037/ser0000075.

Dieterich, William, Christina Mendoza, and Tim Brennan (2016). "COMPAS risk scales: Demonstrating accuracy equity and predictive parity". In: *Northpointe Inc* 7.4.

European Union (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*. https://eur-lex.europa.eu/eli/reg/2016/679/oj.

Flores, Anthony W, Kristin Bechtel, and Christopher T Lowenkamp (2016). "False positives, false negatives, and false analyses: A rejoinder to machine bias: There's

software used across the country to predict future criminals. and it's biased against blacks". In: *Fed. Probation* 80, p. 38.

Hardt, Moritz, Eric Price, and Nati Srebro (2016). "Equality of opportunity in supervised learning". In: *Advances in neural information processing systems* 29.

Kleinberg, Jon M., Sendhil Mullainathan, and Manish Raghavan (2016). "Inherent Trade-Offs in the Fair Determination of Risk Scores". In: *Information Technology Convergence and Services*. URL: https://api.semanticscholar.org/CorpusID:12845273.

Varshney, Kush R (2016). "Engineering safety in machine learning". In: *2016 Information Theory and Applications Workshop (ITA)*. IEEE, pp. 1–5.

Bechavod, Yahav and Katrina Ligett (2017). "Penalizing unfairness in binary classification". In: *arXiv preprint arXiv:1707.00044*.

Berk, Richard, Hoda Heidari, Shahin Jabbari, Matthew Joseph, et al. (2017). "A convex framework for fair regression". In: *arXiv preprint arXiv:1706.02409*.

Beutel, Alex et al. (2017). "Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations". In: *ArXiv* abs/1707.00075. URL: https://api.semanticscholar.org/CorpusID:24990444.

Calmon, F et al. (2017). "Optimized pre-processing for discrimination prevention". In: *Advances in neural information processing systems*.

Chouldechova, Alexandra (2017). "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments". In: *Big data* 5.2, pp. 153–163. DOI: https://doi.org/10.48550/arXiv.1610.07524.

Corbett-Davies, Sam et al. (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.

Doshi-Velez, Finale and Been Kim (2017). "Towards a rigorous science of interpretable machine learning". In: *arXiv preprint arXiv:1702.08608*.

Goodman, Bryce and Seth Flaxman (2017). "European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation"". In: *AI Magazine* 38.3, pp. 50–57.

Hurlburt, George (2017). "How much to trust artificial intelligence?" In: *It Professional* 19.4, pp. 7–11. DOI: https://doi.org/10.1109/MITP.2017.3051326.

Kusner, Matt J et al. (2017). "Counterfactual fairness". In: *Advances in neural information processing systems* 30.

Ozkan, Turgut (2017). "Predicting recidivism through machine learning". PhD thesis.

Pleiss, Geoff et al. (2017). "On fairness and calibration". In: *Advances in neural information processing systems* 30.

Quadrianto, Novi and Viktoriia Sharmanska (2017). "Recycling privileged learning and distribution matching for fairness". In: *Advances in Neural Information Processing Systems* 30.

Voigt, Paul and Axel Von dem Bussche (2017). "The eu general data protection regulation (gdpr)". In: *A practical guide, 1st ed., Cham: Springer International Publishing* 10.3152676, pp. 10–5555.

Zafar, Muhammad Bilal et al. (2017). "Fairness constraints: Mechanisms for fair classification". In: *Artificial intelligence and statistics*. PMLR, pp. 962–970.

Zeng, Jiaming, Berk Ustun, and Cynthia Rudin (2017). "Interpretable classification models for recidivism prediction". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 180.3, pp. 689–722. DOI: http://dx.doi.org/10.1111/rssa.12227.

Zhang, Lu, Yongkai Wu, and Xintao Wu (2017). "A causal framework for discovering and removing direct and indirect discrimination". In: *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pp. 3929–3935.

Agarwal, Alekh et al. (2018). "A reductions approach to fair classification". In: *International conference on machine learning*. PMLR, pp. 60–69.

Binns, Reuben (2018). "Fairness in machine learning: Lessons from political philosophy". In: *Conference on fairness, accountability and transparency*. PMLR, pp. 149–159.

Buolamwini, Joy and Timnit Gebru (2018). "Gender shades: Intersectional accuracy disparities in commercial gender classification". In: *Conference on fairness, accountability and transparency*. PMLR, pp. 77–91.

Chakrabarty, Navoneel and Sanket Biswas (2018). "A statistical approach to adult census income level prediction". In: *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*. IEEE, pp. 207–212.

Douzas, Georgios, Fernando Bacao, and Felix Last (2018). "Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE". In: *Information sciences* 465, pp. 1–20.

Dressel, Julia and Hany Farid (2018). "The accuracy, fairness, and limits of predicting recidivism". In: *Science advances* 4.1, eaao5580.

Eubanks, Virginia (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

Gilpin, Leilani H et al. (2018). "Explaining explanations: An overview of interpretability of machine learning". In: *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, pp. 80–89.

Green, Ben (2018). "Fair" risk assessments: A precarious approach for criminal justice reform". In: *5th Workshop on fairness, accountability, and transparency in machine learning*, pp. 1–5.

Grgic-Hlaca, Nina et al. (2018). "Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction". In: *Proceedings of the 2018 World Wide Web Conference*, pp. 903–912. DOI: https://doi.org/10.48550/arXiv.1802.09548.

Hébert-Johnson, Ursula et al. (2018). "Multicalibration: Calibration for the (computationally-identifiable) masses". In: *International Conference on Machine Learning*. PMLR, pp. 1939–1948.

Kearns, Michael et al. (2018). "Preventing fairness gerrymandering: Auditing and learning for subgroup fairness". In: *International conference on machine learning*. PMLR, pp. 2564–2572.

Krasanakis, Emmanouil et al. (2018). "Adaptive sensitive reweighting to mitigate bias in fairness-aware classification". In: *Proceedings of the 2018 world wide web conference*, pp. 853–862. DOI: https://doi.org/10.1145/3178876.3186133.

Lipton, Zachary C (2018). "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." In: *Queue* 16.3, pp. 31–57. DOI: http://dx.doi.org/10.1145/3233231.

Liu, Lydia T et al. (2018). "Delayed impact of fair machine learning". In: *International Conference on Machine Learning*. PMLR, pp. 3150–3158.

Menon, Aditya Krishna and Robert C Williamson (2018). "The cost of fairness in binary classification". In: *Conference on Fairness, Accountability and Transparency*. PMLR, pp. 107–118.

Pin Calmon, Flavio du et al. (2018). "Data pre-processing for discrimination prevention: Information-theoretic optimization and analysis". In: *IEEE Journal of Selected Topics in Signal Processing* 12.5, pp. 1106–1119. DOI: https://doi.org/10.1109/JSTSP.2018.2865887.

Rajkomar, Alvin et al. (2018). "Ensuring fairness in machine learning to advance health equity". In: *Annals of internal medicine* 169.12, pp. 866–872. DOI: https://doi.org/10.7326/M18-1990.

Reisman, Dillon et al. (2018). "Algorithmic impact assessments: a practical Framework for Public Agency". In: *AI Now* 9.

Veale, Michael, Max Van Kleek, and Reuben Binns (2018). "Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-

making". In: *Proceedings of the 2018 chi conference on human factors in computing systems*, pp. 1–14.

Wadsworth, Christina, Francesca Vera, and Chris Piech (2018). "Achieving fairness through adversarial learning: an application to recidivism prediction". In: *arXiv preprint arXiv:1807.00199*. DOI: https://doi.org/10.48550/arXiv.1807.00199.

Wijenayake, Senuri, Timothy Graham, and Peter Christen (2018). "A decision tree approach to predicting recidivism in domestic violence". In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, pp. 3–15. DOI: https://doi.org/10.48550/arXiv.1803.09862.

Zhang, Brian Hu, Blake Lemoine, and Margaret Mitchell (2018). "Mitigating unwanted biases with adversarial learning". In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340.

AI, HLEG (2019). *High-level expert group on artificial intelligence*.

Amershi, Saleema et al. (May 2019). "Guidelines for Human-AI Interaction". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Glasgow, Scotland: ACM. DOI: https://doi.org/10.1145/3290605.3300233.

Angwin, Julia et al. (2019). "Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks.—-2016". In: URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Bellamy, Rachel KE et al. (2019). "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias". In: *IBM Journal of Research and Development* 63.4/5, pp. 4–1.

Berk, Richard (2019). "Accuracy and fairness for juvenile justice risk assessments". In: *Journal of Empirical Legal Studies* 16.1, pp. 175–194. DOI: https://doi.org/10.1111/jels.12206.

Eckhouse, Laurel et al. (2019). "Layers of bias: A unified approach for understanding problems with risk assessment". In: *Criminal Justice and Behavior* 46.2, pp. 185–209. DOI: https://doi.org/10.1177/0093854818811379.

Al-Faham, Hajer, Angelique M Davis, and Rose Ernst (2019). "Intersectionality: From theory to practice". In: *Annual Review of Law and Social Science* 15.1, pp. 247–265.

Floridi, Luciano (2019). "Establishing the rules for building trustworthy AI". In: *Nature Machine Intelligence* 1.6, pp. 261–262. DOI: https://doi.org/10.1038/s42256-019-0055-y.

Friedler, Sorelle A et al. (2019). "A comparative study of fairness-enhancing interventions in machine learning". In: *Proceedings of the conference on fairness, accountability, and transparency*, pp. 329–338.

Green, Ben and Yiling Chen (2019). "Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments". In: *Proceedings of the conference on fairness, accountability, and transparency*, pp. 90–99. DOI: https://doi.org/10.1145/3287560.3287563.

Gu, Jindong and Daniela Oelke (2019). "Understanding bias in machine learning". In: *arXiv preprint arXiv:1909.01866*.

Hamilton, Melissa (2019). "The sexist algorithm". In: *Behavioral sciences & the law* 37.2, pp. 145–157. DOI: https://doi.org/10.1002/bsl.2406.

Heidari, Hoda and Andreas Krause (2019). "Preventing Disparate Treatment in Sequential Decision Making". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 1. AAAI Press, pp. 2032–2040.

Holstein, Kenneth et al. (2019). "Improving fairness in machine learning systems: What do industry practitioners need?" In: *Proceedings of the 2019 CHI conference on human factors in computing systems*, pp. 1–16. DOI: https://doi.org/10.1145/3290605.3300830.

Iosifidis, Vasileios and Eirini Ntoutsi (2019). "Adafair: Cumulative fairness adaptive boosting". In: *Proceedings of the 28th ACM international conference on infor-*

mation and knowledge management, pp. 781–790. DOI: https://doi.org/10.1145/3357384.3357974.

Jain, Bhanu, Manfred Huber, Leonidas Fegaras, et al. (2019). "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. DOI: https://doi.org/10.1145/3316782.3322787.

Jobin, Anna, Marcello Ienca, and Effy Vayena (2019). "The global landscape of AI ethics guidelines". In: *Nature machine intelligence* 1.9, pp. 389–399. DOI: https://doi.org/10.1038/s42256-019-0088-2.

Keyes, Os, Jevan Hutson, and Meredith Durbin (2019). "A mulching proposal: Analysing and improving an algorithmic system for turning the elderly into high-nutrient slurry". In: *Extended abstracts of the 2019 CHI conference on human factors in computing systems*, pp. 1–11.

Kim, Michael P, Amirata Ghorbani, and James Zou (2019). "Multiaccuracy: Black-box post-processing for fairness in classification". In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 247–254. DOI: https://doi.org/10.1145/3306618.3314287.

Madras, David et al. (2019). "Fairness Through Causal Awareness: Learning Latent-Variable Models for Biased Data". In: *Proceedings of the 2019 ACM Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Atlanta, GA, USA: ACM. DOI: https://doi.org/10.1145/3287560.3287564.

Miller, Tim (2019). "Explanation in artificial intelligence: Insights from the social sciences". In: *Artificial intelligence* 267, pp. 1–38.

Mitchell, Margaret et al. (2019). "Model cards for model reporting". In: *Proceedings of the conference on fairness, accountability, and transparency*, pp. 220–229.

Mittelstadt, Brent (2019). "Principles alone cannot guarantee ethical AI". In: *Nature machine intelligence* 1.11, pp. 501–507.

Montani, Stefania and Manuel Striani (2019). "Artificial intelligence in clinical decision support: a focused literature survey". In: *Yearbook of medical informatics* 28.01, pp. 120–127.

Moriai, Shiho (2019). "Privacy-preserving deep learning via additively homomorphic encryption". In: *2019 IEEE 26th Symposium on Computer Arithmetic (ARITH)*. IEEE Computer Society, pp. 198–198. DOI: https://doi.ieeecomputersociety.org/10.1109/ARITH.2019.00047.

Morina, Giulio et al. (2019). "Auditing and achieving intersectional fairness in classification problems". In: *arXiv preprint arXiv:1911.01468*.

Obermeyer, Ziad et al. (2019). "Dissecting racial bias in an algorithm used to manage the health of populations". In: *Science* 366.6464, pp. 447–453.

OECD (2019). *OECD Principles on Artificial Intelligence*. URL: https://oecd.ai/en/ai-principles.

Raji, Inioluwa Deborah and Joy Buolamwini (2019). "Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products". In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 429–435.

Selbst, Andrew D et al. (2019). "Fairness and abstraction in sociotechnical systems". In: *Proceedings of the conference on fairness, accountability, and transparency*, pp. 59–68.

Shih, Po-Chou, Chui-Yu Chiu, and Chi-Hsun Chou (2019). "Using dynamic adjusting NGHS-ANN for predicting the recidivism rate of commuted prisoners". In: *Mathematics* 7.12, p. 1187. DOI: https://doi.org/10.3390/math7121187.

Suresh, Harini and John V Guttag (2019). "A framework for understanding unintended consequences of machine learning". In: *arXiv preprint arXiv:1901.10002* 2, p. 8.

Sutrop, Margit (2019). "Should we trust artificial intelligence?" In: *Trames: A Journal of the Humanities and Social Sciences* 23.4, pp. 499–522.

Yang, Qiang et al. (2019). "Federated machine learning: Concept and applications". In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 10.2, pp. 1–19. DOI: https://doi.org/10.1145/3298981.

Zelaya, Carlos Vladimiro González (2019). "Towards explaining the effects of data preprocessing on machine learning". In: *2019 IEEE 35th international conference on data engineering (ICDE)*. IEEE, pp. 2086–2090.

Zhao, Han et al. (2019). "Conditional learning of fair representations". In: *arXiv preprint arXiv:1910.07162*.

Beshi, Taye Demissie and Ranvinderjit Kaur (2020). "Public trust in local government: Explaining the role of good governance practices". In: *Public Organization Review* 20.2, pp. 337–350. DOI: https://doi.org/10.1007/s11115-019-00444-6.

Bietti, Elettra (2020). "From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy". In: *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 210–219. DOI: https://doi.org/10.1145/3351095.3372860.

Binns, Reuben (2020). "On the apparent conflict between individual and group fairness". In: *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 514–524. DOI: https://doi.org/10.1145/3351095.3372864.

Birhane, Abeba (2020). "Algorithmic colonization of Africa". In: *SCRIPTed* 17, p. 389. DOI: https://doi.org/10.2966/scrip.170220.389.

Brundage, Miles et al. (2020). "Toward trustworthy AI development: mechanisms for supporting verifiable claims". In: *arXiv preprint arXiv:2004.07213*.

Chakraborty, Joymallya et al. (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. DOI: https://doi.org/10.1145/3368089.3409697.

Costanza-Chock, Sasha (2020). *Design justice: Community-led practices to build the worlds we need*. The MIT Press.

Fjeld, Jessica et al. (2020). "Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI". In: *Berkman Klein Center Research Publication* 2020-1.

Foulds, James R et al. (2020). "An intersectional definition of fairness". In: *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. IEEE, pp. 1918–1921. DOI: https://doi.org/10.1109/ICDE48307.2020.00203.

Green, Ben (2020). "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. DOI: http://doi.org/10.1145/3351095.3372869.

Hanna, Alex et al. (2020a). "Towards a critical race methodology in algorithmic fairness". In: *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 501–512.

— (2020b). "Towards a critical race methodology in algorithmic fairness". In: *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 501–512.

Jain, Bhanu, Manfred Huber, Ramez Elmasri, et al. (2020). "Using bias parity score to find feature-rich models with least relative bias". In: *Technologies* 8.4, p. 68. DOI: https://doi.org/10.3390/technologies8040068.

Jain, Bhanu, Manfred Huber, Ramez A Elmasri, et al. (2020). "Reducing race-based bias and increasing recidivism prediction accuracy by using past criminal history details". In: *Proceedings of the 13th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, pp. 1–8. DOI: https://doi.org/10.1145/3389189.3397990.

Jiang, Heinrich and Ofir Nachum (2020). "Identifying and correcting label bias in machine learning". In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 702–712.

Jurídics i Formació Especialitzada (CEJFE), Centre d'Estudis (2020). *Taxa de reincidència penitenciària 2020*. URL: https://cejfe.gencat.cat/ca/recerca/opendata/presons/taxa-reincidencia-2020/index.html (visited on 04/17/2024).

Khorshidi, Samira, Jeremy G Carter, and George Mohler (2020). "Repurposing recidivism models for forecasting police officer use of force". In: *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, pp. 3199–3203. DOI: https://doi.org/10.1109/BigData50022.2020.9378173.

Kilbertus, Niki et al. (2020). "Fair decisions despite imperfect predictions". In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 277–287.

Kobayashi, Kenji and Yuri Nakao (2020). "One-vs.-One Mitigation of Intersectional Bias: A General Method to Extend Fairness-Aware Binary Classification". In: *arXiv preprint arXiv:2010.13494*.

Lo Piano, Samuele (2020). "Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward". In: *Humanities and Social Sciences Communications* 7.1, pp. 1–7. DOI: https://doi.org/10.1057/s41599-020-0501-9.

McKay, Carolyn (2020). "Predicting risk in criminal procedure: actuarial tools, algorithms, AI and judicial decision-making". In: *Current Issues in Criminal Justice* 32.1, pp. 22–39. DOI: https://doi.org/10.1080/10345329.2019.1658694.

Mohammed, Roweida, Jumanah Rawashdeh, and Malak Abdullah (2020). "Machine learning with oversampling and undersampling techniques: overview study and experimental results". In: *2020 11th international conference on information and communication systems (ICICS)*. IEEE, pp. 243–248.

O'Donnell, Ian (2020). "An evidence review of recidivism and policy responses". In: *Department of Justice & Equality, Dublin*.

Raji, Inioluwa Deborah, Andrew Smart, et al. (2020). "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing". In:

Proceedings of the 2020 conference on fairness, accountability, and transparency, pp. 33–44. DOI: https://doi.org/10.1145/3351095.3372873.

Ribeiro, Marco Tulio et al. (2020). "Beyond Accuracy: Behavioral Testing of NLP Models with CheckList". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 4902–4912. DOI: https://doi.org/10.18653/v1/2020.acl-main.442.

Rodolfa, Kit T et al. (2020). "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. DOI: https://doi.org/10.1145/3351095.3372863.

Rudin, Cynthia, Caroline Wang, and Beau Coker (2020). "The age of secrecy and unfairness in recidivism prediction". In: *Harvard Data Science Review* 2.1, p. 1. DOI: https://doi.org/10.48550/arXiv.1811.00731.

Ryan, Mark (2020). "In AI we trust: ethics, artificial intelligence, and reliability". In: *Science and Engineering Ethics* 26.5, pp. 2749–2767. DOI: https://doi.org/10.1007/s11948-020-00228-y.

Schwind, Nicolas et al. (2020). "Representative solutions for bi-objective optimisation". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34, pp. 1436–1443.

Skeem, Jennifer and Christopher Lowenkamp (2020). "Using algorithms to address trade-offs inherent in predicting recidivism". In: *Behavioral Sciences & the Law* 38.3, pp. 259–278. DOI: https://doi.org/10.1002/bsl.2465.

Stoyanovich, Julia, Bill Howe, and Hosagrahar Visvesvaraya Jagadish (2020). "Responsible data management". In: *Proceedings of the VLDB Endowment* 13.12.

Strubell, Emma, Ananya Ganesh, and Andrew McCallum (2020). "Energy and policy considerations for modern deep learning research". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. 09, pp. 13693–13696.

Sushina, Tatyana and Andrew Sobenin (2020). "Artificial Intelligence in the Criminal Justice System: Leading Trends and Possibilities". In: *6th International Con-*

ference on Social, economic, and academic leadership (ICSEAL-6-2019). Atlantis Press, pp. 432–437. DOI: https://doi.org/10.2991/assehr.k.200526.062.

Toreini, Ehsan et al. (2020). "The relationship between trust in AI and trustworthy machine learning technologies". In: *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 272–283. DOI: https://doi.org/10.48550/arXiv.1912.00782.

Vinuesa, Ricardo et al. (2020). "The role of artificial intelligence in achieving the Sustainable Development Goals". In: *Nature communications* 11.1, p. 233. DOI: https://doi.org/10.1038/s41467-019-14108-y.

Xu, Zhaozhao et al. (2020). "A hybrid sampling algorithm combining M-SMOTE and ENN based on Random forest for medical imbalanced data". In: *Journal of Biomedical Informatics* 107, p. 103465.

Yeung, Karen (2020). "Recommendation of the council on artificial intelligence (OECD)". In: *International legal materials* 59.1, pp. 27–34.

Alikhademi, Kiana et al. (2021). "A review of predictive policing from the perspective of fairness". In: *Artificial Intelligence and Law* 7, pp. 1–17. DOI: https://doi.org/10.1007/s10506-021-09286-4.

Barda, Noam et al. (2021). "Addressing bias in prediction models by improving subpopulation calibration". In: *Journal of the American Medical Informatics Association* 28.3, pp. 549–558. DOI: https://doi.org/10.1093/jamia/ocaa283.

Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns, et al. (2021). "Fairness in criminal justice risk assessments: The state of the art". In: *Sociological Methods & Research* 50.1, pp. 3–44. DOI: https://doi.org/10.1177/0049124118782533.

Birhane, Abeba (2021). "Algorithmic injustice: a relational ethics approach". In: *Patterns* 2.2.

Biswas, Arpita and Suvam Mukherjee (2021). "Ensuring fairness under prior probability shifts". In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics,*

and Society. New York, NY, USA: Association for Computing Machinery, pp. 414–424. DOI: https://doi.org/10.1145/3461702.3462596.

Bommasani, Rishi (2021). "On the opportunities and risks of foundation models". In: *arXiv preprint arXiv:2108.07258*.

Chugh, Neha (2021). "Risk assessment tools on trial: Lessons learned for "Ethical AI" in the criminal justice system". In: *2021 IEEE International Symposium on Technology and Society (ISTAS)*. Waterloo, ON, Canada: IEEE, pp. 1–5. DOI: https://doi.org/10.1109/ISTAS52410.2021.9629143.

Crawford, Kate (2021). *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven, CT: Yale University Press. ISBN: 9780300264630. DOI: https://doi.org/10.5465/amle.2025.0053.

Floridi, Luciano (2021). "The European legislation on AI: a brief analysis of its philosophical approach". In: *Philosophy & Technology* 34.2, pp. 215–222. DOI: https://doi.org/10.1007/s13347-021-00460-9.

Gebru, Timnit et al. (2021). "Datasheets for datasets". In: *Communications of the ACM* 64.12, pp. 86–92. DOI: http://dx.doi.org/10.1145/3458723.

Ghosh, Avijit, Lea Genuit, and Mary Reagan (2021). "Characterizing intersectional group fairness with worst-case comparisons". In: *Artificial Intelligence Diversity, Belonging, Equity, and Inclusion*. PMLR, pp. 22–34.

Hartmann, Kathrin and Georg Wenzelburger (2021). "Uncertainty, risk and the use of algorithms in policy decisions: a case study on criminal justice in the USA". In: *Policy Sciences* 54.2, pp. 269–287. DOI: https://doi.org/10.1007/s11077-020-09414-y.

Karimi-Haghighi, Marzieh and Carlos Castillo (2021a). "Efficiency and fairness in recurring data-driven risk assessments of violent recidivism". In: *Proceedings of the 36th annual acm symposium on applied computing*, pp. 994–1002. DOI: https://doi.org/10.1145/3412841.3441975.

— (2021b). "Enhancing a recidivism prediction tool with machine learning: effectiveness and algorithmic fairness". In: *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pp. 210–214.

Kasy, Maximilian and Rediet Abebe (2021). "Fairness, equality, and power in algorithmic decision-making". In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 576–586.

Liu, Jiashuo et al. (2021). "Towards out-of-distribution generalization: A survey". In: *arXiv preprint arXiv:2108.13624*.

Mehrabi, Ninareh et al. (2021). "A survey on bias and fairness in machine learning". In: *ACM Computing Surveys (CSUR)* 54.6, pp. 1–35.

Miron, Marius et al. (2021). "Evaluating causes of algorithmic bias in juvenile criminal recidivism". In: *Artificial Intelligence and Law* 29.2, pp. 111–147. DOI: https://doi.org/10.1007/s10506-020-09268-y.

Mishler, Alan, Edward H Kennedy, and Alexandra Chouldechova (2021). "Fairness in risk assessment instruments: Post-processing to achieve counterfactual equalized odds". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 386–400.

Mohler, George and Michael D Porter (2021). "A note on the multiplicative fairness score in the NIJ recidivism forecasting challenge". In: *Crime Science* 10.1, pp. 1–5. DOI: https://doi.org/10.1186/s40163-021-00152-x.

O'Loughlin, Timothy and Rachel Bukowitz (2021). "A new approach toward social licensing of data analytics in the public sector". In: *Australian Journal of Social Issues* 56.2, pp. 198–212. DOI: https://doi.org/10.1002/ajs4.161.

OECD (2021). "Tools for trustworthy AI: A framework to compare implementation tools for trustworthy AI systems". In: 312. DOI: https://doi.org/10.1787/008232ec-en. URL: https://www.oecd-ilibrary.org/content/paper/008232ec-en.

Rančić, Sanja, Sandro Radovanović, and Boris Delibašić (2021). "Investigating oversampling techniques for fair machine learning models". In: *Decision Support Sys-*

tems XI: Decision Support Systems, Analytics and Technologies in Response to Global Crisis Management: 7th International Conference on Decision Support System Technology, ICDSST 2021, Loughborough, UK, May 26–28, 2021, Proceedings. Springer, pp. 110–123.

Roh, Yuji et al. (2021). "Fairbatch: Batch selection for model fairness". In: *arXiv preprint arXiv:2012.01696*.

Salazar, Teresa et al. (2021). "Fawos: Fairness-aware oversampling algorithm based on distributions of sensitive attributes". In: *IEEE Access* 9, pp. 81370–81379.

Standardization, International Organization for (2021). *Information technology – Artificial Intelligence (AI) – Bias in AI systems and AI aided decision making*. Standard ISO/IEC TR 24027:2021(E). Vernier, Geneva, Switzerland: International Organization for Standardization. URL: `https://www.iso.org/standard/77607.html`.

Thiebes, Scott, Sebastian Lins, and Ali Sunyaev (2021). "Trustworthy artificial intelligence". In: *Electronic Markets* 31.2, pp. 447–464.

UNESCO (2021). *Ethics of Artificial Intelligence*. URL: `https://www.unesco.org/en/artificial-intelligence/recommendation-ethics`.

Wang, Jia-Bao, Chun-An Zou, and Guang-Hui Fu (2021). "AWSMOTE: An SVM-Based Adaptive Weighted SMOTE for Class-Imbalance Learning". In: *Scientific Programming* 2021.1, p. 9947621.

Wen, Min, Osbert Bastani, and Ufuk Topcu (2021). "Algorithmic Fairness in Sequential Decision Making: Stability, Fairness, and Performance". In: *Proceedings of the 35th AAAI Conference on Artificial Intelligence*. AAAI Press, pp. 9602–9611.

Angwin, Julia et al. (2022). "Machine bias". In: *Ethics of data and analytics*. Auerbach Publications, pp. 254–264.

Bayram, Firas, Bestoun S Ahmed, and Andreas Kassler (2022). "From concept drift to model degradation: An overview on performance-aware drift detectors". In:

Knowledge-Based Systems 245, p. 108632. DOI: https://doi.org/10.1016/j.knosys.2022.108632.

Castelnovo, Alessandro et al. (2022). "FFTree: A flexible tree to handle multiple fairness criteria". In: *Information Processing & Management* 59.6, p. 103099.

Choudhary, Vishwas et al. (2022). "Detecting Concept Drift in the Presence of Sparsity–A Case Study of Automated Change Risk Assessment System". In: *arXiv preprint arXiv:2207.13287*.

Crenshaw, Kimberlé (2022). "Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics [1989]". In: *Contemporary sociological theory* 1, p. 354.

Dass, Rahul Kumar et al. (Mar. 2022). "Detecting racial inequalities in criminal justice: towards an equitable deep learning approach for generating and interpreting racial categories using mugshots". In: *AI & SOCIETY*, pp. 1–22. DOI: https://doi.org/10.1007/s00146-022-01440-z.

Figueroa-Armijos, Maria, Brent B Clark, and Serge P da Motta Veiga (2022). "Ethical perceptions of AI in hiring and organizational trust: The role of performance expectancy and social influence". In: *Journal of Business Ethics*, pp. 1–19. DOI: https://doi.org/10.1007/s10551-022-05166-2.

Gao, Xuanqi et al. (2022). "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. DOI: https://doi.org/10.1145/3510003.3510087.

Goodman, Ellen P and Julia Trehu (2022). "Algorithmic auditing: Chasing AI accountability". In: *Santa Clara High Tech. LJ* 39, p. 289.

Kaur, Davinder et al. (2022). "Trustworthy artificial intelligence: a review". In: *ACM Computing Surveys (CSUR)* 55.2, pp. 1–38.

Kokhlikyan, Narine et al. (2022). "Bias mitigation framework for intersectional subgroups in neural networks". In: *arXiv preprint arXiv:2212.13014*.

Kozodoi, Nikita, Johannes Jacob, and Stefan Lessmann (2022). "Fairness in credit scoring: Assessment, implementation and profit implications". In: *European Journal of Operational Research* 297.3, pp. 1083–1094.

Le Quy, Tai et al. (2022). "A survey on datasets for fairness-aware machine learning". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 12.3, e1452.

Lewis, Grace A et al. (2022). "Augur: A step towards realistic drift detection in production ml systems". In: *Proceedings of the 1st Workshop on Software Engineering for Responsible AI*, pp. 37–44. DOI: https://doi.org/10.1145/3526073.3527590.

Liu, Haochen et al. (2022). "Trustworthy ai: A computational perspective". In: *ACM Transactions on Intelligent Systems and Technology* 14.1, pp. 1–59. DOI: https://doi.org/10.1145/3546872.

Liu, Kaifeng and Da Tao (2022). "The roles of trust, personalization, loss of privacy, and anthropomorphism in public acceptance of smart healthcare services". In: *Computers in Human Behavior* 127, p. 107026. DOI: https://doi.org/10.1016/j.chb.2021.107026.

Liu, Suyun and Luis Nunes Vicente (2022). "Accuracy and fairness trade-offs in machine learning: A stochastic multi-objective approach". In: *Computational Management Science* 19.3, pp. 513–537. DOI: https://doi.org/10.1007/s10287-022-00425-z.

Mökander, Jakob et al. (2022). "The US Algorithmic Accountability Act of 2022 vs. The EU Artificial Intelligence Act: what can they learn from each other?" In: *Minds and Machines*, pp. 1–8. DOI: https://doi.org/10.1007/s11023-022-09612-y.

Oatley, Giles C (2022). "Themes in data mining, big data, and crime analytics". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 12.2, e1432. DOI: https://doi.org/10.1002/widm.1432.

Pessach, Dana and Erez Shmueli (2022). "A review on fairness in machine learning". In: *ACM Computing Surveys (CSUR)* 55.3, pp. 1–44.

Ratih, Iis Dewi et al. (2022). "Synthetic minority over-sampling technique nominal continous logistic regression for imbalanced data". In: *AIP Conference Proceedings*. Vol. 2668. 1. AIP Publishing.

Sharma, Shubhkirti and Vijay Kumar (2022). "A comprehensive review on multi-objective optimization techniques: Past, present and future". In: *Archives of Computational Methods in Engineering* 29.7, pp. 5605–5633.

Varona, Daniel and Juan Luis Suárez (2022). "Discrimination, Bias, Fairness, and Trustworthy AI". In: *Applied Sciences* 12.12, p. 5826.

Vela, Daniel et al. (2022). "Temporal quality degradation in AI models". In: *Scientific reports* 12.1, p. 11654. DOI: https://doi.org/10.1038/s41598-022-15245-z.

Wang, Angelina, Vikram V Ramaswamy, and Olga Russakovsky (2022). "Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation". In: *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pp. 336–349.

Wang, Caroline et al. (2022). "In pursuit of interpretable, fair and accurate machine learning for criminal recidivism prediction". In: *Journal of Quantitative Criminology*, pp. 1–63. DOI: https://10.1007/s10940-022-09545-w.

Weidinger, Laura et al. (2022). "Taxonomy of risks posed by language models". In: *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pp. 214–229.

White House Office of Science and Technology Policy (2022). *Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People*. Tech. rep. The White House. URL: https://www.whitehouse.gov/ostp/ai-bill-of-rights.

Yao, Huaxiu et al. (2022). "Wild-time: A benchmark of in-the-wild distribution shift over time". In: *Advances in Neural Information Processing Systems* 35, pp. 10309–10324.

Zeng, Xianli, Edgar Dobriban, and Guang Cheng (2022). "Fair Bayes-Optimal Classifiers Under Predictive Parity". In: *arXiv preprint arXiv:2205.07182*.

Zhang, Mengdi and Jun Sun (2022). "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. DOI: https://doi.org/10.1145/3540250.3549103.

Ződi, Zsolt (2022). "Algorithmic explainability and legal reasoning". In: *The Theory and Practice of Legislation* 10.1, pp. 67–92. DOI: https://doi.org/10.1080/20508840.2022.2033945.

AI, NIST (2023). "Artificial intelligence risk management framework (AI RMF 1.0)". In: pp. 100–1. DOI: https://doi.org/10.6028/NIST.AI.100-1.

Barocas, Solon, Moritz Hardt, and Arvind Narayanan (2023). *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.

Bonta, James and Donald Arthur Andrews (2023). *The psychology of criminal conduct*. Routledge.

Cai, Tiffany Tianhui, Hongseok Namkoong, and Steve Yadlowsky (2023). "Diagnosing model performance under distribution shift". In: *arXiv preprint arXiv:2303.02011*.

Caton, Simon and Christian Haas (2023). "Fairness in Machine Learning: A Survey". In: *ACM Comput. Surv.*, pp. 360–0300. DOI: https://doi.org/10.1145/3616865.

Chen, Zhenpeng et al. (2023). "A comprehensive empirical study of bias mitigation methods for machine learning classifiers". In: *ACM transactions on software engineering and methodology* 32.4, pp. 1–30.

Chiou, Erin K and John D Lee (2023). "Trusting automation: Designing for responsivity and resilience". In: *Human factors* 65.1, pp. 137–165.

Delaney, Eoin et al. (2023). "Counterfactual explanations for misclassified images: How human and machine explanations differ". In: *Artificial Intelligence* 324, p. 103995. DOI: https://doi.org/10.1016/j.artint.2023.103995.

Delgado, Fernando et al. (2023). "The participatory turn in ai design: Theoretical foundations and the current state of practice". In: *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pp. 1–23. DOI: https://doi.org/10.1145/3617694.3623261.

Farayola, Michael Mayowa, Irina Tal, Regina Connolly, et al. (2023). "Ethics and Trustworthiness of AI for Predicting the Risk of Recidivism: A Systematic Literature Review". In: *Information* 14.8, p. 426. DOI: https://doi.org/10.3390/info14080426.

Farayola, Michael Mayowa, Irina Tal, Bendechache Malika, et al. (2023). "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. DOI: https://doi.org/10.1145/3600160.3605033.

Ghnemat, Rawan, Sawsan Alodibat, and Qasem Abu Al-Haija (2023). "Explainable artificial intelligence (XAI) for deep learning based medical imaging classification". In: *Journal of Imaging* 9.9, p. 177. DOI: https://doi.org/10.3390/jimaging9090177.

Gohar, Usman and Lu Cheng (2023). "A survey on intersectional fairness in machine learning: Notions, mitigation, and challenges". In: *arXiv preprint arXiv:2305.06969*.

Hickman, Louis et al. (2023). "Oversampling Higher-Performing Minorities During Machine Learning Model Training Reduces Adverse Impact Slightly but Also Reduces Model Accuracy". In: *arXiv preprint arXiv:2304.13933*.

International Organization for Standardization (2023). *ISO/IEC 42001:2023 Artificial intelligence — Management system*. URL: https://www.iso.org/standard/81230.html.

Islam, Rashidul et al. (2023). "Differential fairness: an intersectional framework for fair AI". In: *Entropy* 25.4, p. 660.

Kim, Savina et al. (2023). "Fair models in credit: Intersectional discrimination and the amplification of inequity". In: *arXiv preprint arXiv:2308.02680*.

Mhlambi, Sábëlo and Simona Tiribelli (2023). "Decolonizing AI ethics: Relational autonomy as a means to counter AI harms". In: *Topoi* 42.3, pp. 867–880. DOI: https://doi.org/10.1007/s11245-022-09874-2.

National Institute of Standards and Technology (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. Tech. rep. U.S. Department of Commerce. DOI: https://doi.org/10.6028/NIST.AI.100-1. URL: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

Ovalle, Anaelia et al. (2023). "Factoring the matrix of domination: A critical review and reimagination of intersectionality in ai fairness". In: *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 496–511. DOI: https://doi.org/10.1145/3600211.3604705.

Pelegrina, Guilherme Dean, Miguel Couceiro, and Leonardo Tomazeli Duarte (2023). "A statistical approach to detect sensitive features in a group fairness setting". In: *arXiv preprint arXiv:2305.06994*.

Raza, Shaina, Parisa Osivand Pour, and Syed Raza Bashir (2023). "Fairness in machine learning meets with equity in healthcare". In: *Proceedings of the AAAI Symposium Series*, pp. 149–153.

Roy, Arjun, Jan Horstmann, and Eirini Ntoutsi (2023). "Multi-dimensional discrimination in law and machine learning-A comparative overview". In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 89–100. DOI: https://doi.org/10.1145/3593013.3593979.

Sonoda, Ryosuke (2023). "Fair oversampling technique using heterogeneous clusters". In: *Information Sciences* 640, p. 119059. DOI: https://doi.org/10.1016/j.ins.2023.119059.

UNESCO (Nov. 2023). *Leveraging UNESCO Normative Instruments for an Ethical Generative AI Use of Indigenous Data*. https://www.unesco.org/en/articles/leveraging-unesco-normative-instruments-ethical-generative-ai-use-indigenous-data. Last updated: 22 March 2024; Accessed: 2025-09-12.

Wan, Mingyang et al. (2023). "In-processing modeling techniques for machine learning fairness: A survey". In: *ACM Transactions on Knowledge Discovery from Data* 17.3, pp. 1–27.

Zhou, Yan, Murat Kantarcioglu, and Chris Clifton (2023). "On improving fairness of AI models with synthetic minority oversampling techniques". In: *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*. SIAM, pp. 874–882.

Adler, Julian, J Antoine, and L Al-Saadoon (2024). "Minding the machines: On values and AI in the criminal legal space". In: *Center for Justice Innovation*.

Berk, Richard A, Arun Kumar Kuchibhotla, and Eric Tchetgen Tchetgen (2024). "Improving fairness in criminal justice algorithmic risk assessments using optimal transport and conformal prediction sets". In: *Sociological Methods & Research* 53.4, pp. 1629–1675. DOI: https://doi.org/10.1177/00491241231155883.

Caton, Simon and Christian Haas (2024). "Fairness in machine learning: A survey". In: *ACM Computing Surveys* 56.7, pp. 1–38.

Christian, Gideon (2024). "Legal Framework for the Use of Artificial Intelligence (AI) Technology in the Canadian Criminal Justice System". In: *Canadian Journal of Law and Technology* 21.2, p. 109.

Cofone, Ignacio and Warut Khern-am-nuai (2024). "The Overstated Cost of AI Fairness in Criminal Justice". In: *Ind. LJ* 100, p. 1431.

Di Gennaro, Federico et al. (2024). "Post-processing fairness with minimal changes". In: *arXiv preprint arXiv:2408.15096*.

European Parliament and Council (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council*. Official Journal of the European Union, L 135, 1–121. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

Farayola, Michael Mayowa, Malika Bendechache, Takfarinas Saber, et al. (2024a). "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.

— (2024b). "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.

Farayola, Michael Mayowa, Tal Irina, et al. (2024). "Towards Trustworthy AI: Potential and Peril of Integrating Multi-Phase Bias Mitigation Techniques in Recidivism Models". In: *Artificial Intelligence*, p. 104394.

Fazel, Seena et al. (2024). "An updated evidence synthesis on the Risk-Need-Responsivity (RNR) model: Umbrella review and commentary". In: *Journal of Criminal Justice* 92, p. 102197. DOI: https://doi.org/10.1016/j.jcrimjus.2024.102197.

Finocchiaro, Jessie (2024). "Using Property Elicitation to Understand the Impacts of Fairness Regularizers". In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 62–73.

Gundhus, Helene OI (2024). "Selective navigation: risk prediction cultures in criminal justice practices". In: *Justice, Power and Resistance* 7.2, pp. 111–128. DOI: https://doi.org/10.1332/26352338Y2024D000000021.

Kabir, Md Alamgir et al. (2024). "Balancing fairness: unveiling the potential of SMOTE-driven oversampling in AI model enhancement". In: *Proceedings of the 2024 9th International Conference on Machine Learning Technologies*, pp. 21–29. DOI: https://doi.org/10.1145/3674029.3674034.

Khreisat, Mohammad N et al. (2024). "Ethical implications of AI integration in educational decision making: Systematic review". In: *Educational Administration: Theory and Practice* 30.5, pp. 8521–8527.

Machado, Agathe Fernandes et al. (2024). "From Uncertainty to Precision: Enhancing Binary Classifier Performance through Calibration". In: *arXiv preprint arXiv:2402.07790*.

McCormack, Louise and Malika Bendechache (2024). "Ethical ai governance: Methods for evaluating trustworthy ai". In: *arXiv preprint arXiv:2409.07473*.

Murch, W Spencer, Sylvia Kairouz, and Martin French (2024). "Establishing the temporal stability of machine learning models that detect online gambling-related harms". In: *Computers in Human Behavior Reports* 14, p. 100427. DOI: https://doi.org/10.1016/j.chbr.2024.100427.

Nagpal, Rashmi et al. (2024). "A multi-objective framework for balancing fairness and accuracy in debiasing machine learning models". In: *Machine Learning and Knowledge Extraction* 6.3, pp. 2130–2148. DOI: https://doi.org/10.3390/make6030105.

Ni, Hongliang et al. (2024). "Fairness without sensitive attributes via knowledge sharing". In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1897–1906. DOI: https://doi.org/10.1145/3630106.3659014.

Pinkava, Thomas, Jack McFarland, and Afra Mashhadi (2024). "A model-and data-agnostic debiasing system for achieving equalized odds". In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. Vol. 7. 1, pp. 1123–1131.

Plecko, Drago and Elias Bareinboim (2024). "Fairness-Accuracy Trade-Offs: A Causal Perspective". In: *arXiv preprint arXiv:2405.15443*.

Popoola, Gideon and John Sheppard (2024). "Investigating and Mitigating the Performance–Fairness Tradeoff via Protected-Category Sampling". In: *Electronics* 13.15, p. 3024. DOI: https://doi.org/10.3390/electronics13153024.

Ramachandranpillai, Resmi et al. (2024). "Fairness at every intersection: Uncovering and mitigating intersectional biases in multimodal clinical predictions". In: *arXiv preprint arXiv:2412.00606*.

Räz, Tim (2024). "Reliability Gaps Between Groups in COMPAS Dataset". In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 113–126. DOI: https://doi.org/10.1145/3630106.3658544.

Ren, Shaolei and Adam Wierman (2024). *The uneven distribution of AI's environmental impacts*. Harvard Business Review.

Scaria, Arul George et al. (2024). "Algorithms and recidivism: A multi-disciplinary systematic review". In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. Vol. 7, pp. 1292–1305.

Schrouff, Jessica et al. (2024). "Mind the graph when balancing data for fairness or robustness". In: *Advances in Neural Information Processing Systems* 37, pp. 29913–29947.

Sheppard, Keller G, Alyssa R Talaugon, and Jorge L Hernandez (2024). "Assessing the feasibility and performance of risk assessment instruments for early intervention and prevention services in Juvenile Justice". In: *Journal of Criminal Justice* 94, p. 102262. DOI: https://doi.org/10.1016/j.jcrimjus.2024.102262.

Sikder, Md Fahim, Daniel de Leng, and Fredrik Heintz (2024). "Fair4Free: Generating High-fidelity Fair Synthetic Samples using Data Free Distillation". In: *arXiv preprint arXiv:2410.01423*.

Small, Edward et al. (2024). "Equalised odds is not equal individual odds: Post-processing for group and individual fairness". In: *Proceedings of the 2024 ACM conference on Fairness, Accountability, and Transparency*, pp. 1559–1578.

Sousa, Sonia et al. (2024). "Human-centered trustworthy framework: A human–computer interaction perspective". In: *Computer* 57.3, pp. 46–58.

U.S. Department of Justice (2024). *Artificial Intelligence and Criminal Justice: Final Report*. Tech. rep. U.S. Department of Justice. URL: https://www.justice.gov/olp/media/1381796/dl.

Valentine, Alissa A, Alexander W Charney, and Isotta Landi (2024). "Fair Machine Learning for Healthcare Requires Recognizing the Intersectionality of Sociodemographic Factors, a Case Study". In: *arXiv preprint arXiv:2407.15006*.

Wang, Chuqiao, Junru Li, and Ruiming Zhang (2024). "A method to enhance structural fairness in large language models with active learning". In.

Yang, Jingkang et al. (2024). "Generalized out-of-distribution detection: A survey". In: *International Journal of Computer Vision* 132.12, pp. 5635–5662.

Yu, Zhe, Joymallya Chakraborty, and Tim Menzies (2024). "Fairbalance: How to achieve equalized odds with data pre-processing". In: *IEEE Transactions on Software Engineering*.

Zanna, Khadija and Akane Sano (2024). "Enhancing Fairness and Performance in Machine Learning Models: A Multi-Task Learning Approach with Monte-Carlo Dropout and Pareto Optimality". In: *arXiv preprint arXiv:2404.08230*.

Zhao, Han (2024). "Fair and optimal prediction via post-processing". In: *AI Magazine* 45.3, pp. 411–418.

Andrews, Kenya S (2025). "Moving from Fairness to Justice: Intentional Algorithmic Solutions Through an Intersectional Lens". In: *Interactions* 32.5, pp. 58–60. DOI: https://doi.org/10.1145/3760549.

Bonta, James and Seung C Lee (2025). "The Risk–Need–Responsivity Model and Justice-Involved Persons with Serious Mental Illness". In: *Canadian Journal of Criminology and Criminal Justice* 67.1, pp. 88–108. DOI: http://dx.doi.org/10.3138/cjccj-2025-0003.

California Department of Corrections and Rehabilitation (2025). *Latest CDCR Recidivism Report Highlights Decline in Recidivism and Value of Rehabilitative Programming*. URL: https://www.cdcr.ca.gov/news/2025/04/02/latest-cdcr-recidivism-report-highlights-decline-in-recidivism-and-value-of-rehabilitative-programming/.

Carvalho, Miguel, Armando J Pinho, and Susana Brás (2025). "Resampling approaches to handle class imbalance: a review from a data perspective". In: *Journal of Big Data* 12.1, p. 71. DOI: https://doi.org/10.1186/s40537-025-01119-4.

Cavus, Muhammed et al. (2025). "Transparent and bias-resilient AI framework for recidivism prediction using deep learning and clustering techniques in criminal

justice". In: *Applied Soft Computing*, p. 113160. DOI: https://doi.org/10.1016/j.asoc.2025.113160.

Duwe, Grant and Valerie Clark (2025). "The Gap Between the Ideal and Reality in Risk-Needs-Responsivity Assessments: A Survey of US Prisons". In: *Corrections*, pp. 1–18. DOI: https://doi.org/10.1080/23774657.2025.2533132.

Eberhard, Léandre et al. (2025). "General Post-Processing Framework for Fairness Adjustment of Machine Learning Models". In: *arXiv preprint arXiv:2504.16238*.

Engel, Christoph, Lorenz Linhardt, and Marcel Schubert (2025). "Code is law: how COMPAS affects the way the judiciary handles the risk of recidivism: C. Engel et al." In: *Artificial Intelligence and Law* 33.2, pp. 383–404. DOI: https://doi.org/10.1007/s10506-024-09389-8.

Fabris, Alessandro et al. (2025). "Fairness and bias in algorithmic hiring: A multidisciplinary survey". In: *ACM Transactions on Intelligent Systems and Technology* 16.1, pp. 1–54.

Farayola, Michael et al. (2025). "Beyond Aggregate Fairness: Intersectional Auditing Across the AI Fairness Pipeline". In: *AI and Ethics*.

Farayola, Michael Mayowa, Malika Bendechache, Saber Takfarinas, et al. (2025). "Investigating Fairness-Aware Oversampling Strategies and Techniques Across Diverse Machine Learning Algorithms for Recidivism Prediction". In: *Minds and Machines* 35.3, p. 37.

Farayola, Michael Mayowa, Shane Kennedy, et al. (2025). "Intersectional Fairness in Healthcare AI: A Pipeline-Wide Evaluation of Multi-Stage Mitigation Strategies". In.

Farayola, Michael Mayowa, Irina Tal, Takfarinas Saber, et al. (2025). "A fairness-focused approach to recidivism prediction: implications for accuracy, trust, and equity". In: *AI & SOCIETY*, pp. 1–19. DOI: https://doi.org/10.1007/s00146-025-02452-1.

Han, Jessy Xinyi, Kristjan H. Greenewald, and Devavrat Shah (2025). "Fairness Is More Than Algorithms: Racial Disparities in Time-to-Recidivism". In: *ArXiv* abs/2504.18629. URL: https://api.semanticscholar.org/CorpusID:278165729.

Hu, Lily (2025). "Does calibration mean what they say it means; or, the reference class problem rises again". In: *Philosophical Studies*, pp. 1–27.

Inocêncio Júnior, Ronaldo et al. (2025). "Data Balancing for Mitigating Sampling Bias in Machine Learning". In: *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing*, pp. 1204–1212. DOI: https://doi.org/10.1145/3672608.3707891.

Jegham, Nidhal et al. (2025). "How Hungry is AI? Benchmarking Energy, Water, and Carbon Footprint of LLM Inference". In: *ArXiv* abs/2505.09598. DOI: https://doi.org/10.48550/arXiv.2505.09598. URL: https://api.semanticscholar.org/CorpusID:278602401.

Joseph, Jeena (2025). "Predicting crime or perpetuating bias? The AI dilemma". In: *AI & SOCIETY* 40.4, pp. 2319–2321. DOI: https://doi.org/10.1007/s00146-024-02032-9.

Jung, Sandy et al. (2025). "Criminogenic and Non-Criminogenic Factors and Their Association With Reintegration Success for Individuals Under Judicial Orders in Canada". In: *International Journal of Offender Therapy and Comparative Criminology* 69.12, pp. 1688–1706. DOI: https://doi.org/10.1177/0306624X241270603.

Kinney, David (2025). "Aggregating Measures of Accuracy and Fairness in Prediction Algorithms". In: DOI: https://doi.org/10.1145/3715275.3732031.

Koçak, Burak et al. (2025). "Bias in artificial intelligence for medical imaging: fundamentals, detection, avoidance, mitigation, challenges, ethics, and prospects". In: *Diagnostic and interventional radiology* 31.2, p. 75. DOI: https://doi.org/10.4274/dir.2024.242854.

Kristofik, Andrej (2025). "Bias in AI (Supported) Decision Making: Old Problems, New Technologies". In: *IJCA*. Vol. 16. HeinOnline, p. 1.

Laakom, Firas, Haobo Chen, Jurgen Schmidhuber, et al. (2025). "Fairness Overfitting in Machine Learning: An Information-Theoretic Perspective". In: *ArXiv* abs/2506.07861. DOI: https://doi.org/10.48550/arXiv.2507.05823. URL: https://api.semanticscholar.org/CorpusID:279250747.

Laakom, Firas, Haobo Chen, Jürgen Schmidhuber, et al. (2025). "Fairness Overfitting in Machine Learning: An Information-Theoretic Perspective". In: *Forty-second International Conference on Machine Learning*. DOI: https://doi.org/10.48550/arXiv.2506.07861. URL: https://openreview.net/forum?id=saBXnGIDSj.

Law Commission of Ontario (2025). *AI and the Assessment of Risk in Bail, Sentencing, and Parole Decisions*. Tech. rep. Law Commission of Ontario. URL: https://www.lco-cdo.org/wp-content/uploads/2025/04/LCO-AI-in-Criminal-Justice-Paper-3-AI-and-Risk-Assessment.pdf.

Lett, Elle et al. (2025). "Intersectional and Marginal Debiasing in Prediction Models for Emergency Admissions". In: *JAMA Network Open* 8.5, e2512947–e2512947. DOI: https://doi.org/10.1001/jamanetworkopen.2025.12947.

Liu, Qinyi et al. (2025). "Can synthetic data be fair and private? A comparative study of synthetic data generation and fairness algorithms". In: *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, pp. 591–600.

Magaña, Maria Isabel and Katie Shilton (2025). "Frameworks, Methods and Shared Tasks: Connecting Participatory AI to Trustworthy AI Through a Systematic Review of Global Projects". In: *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2166–2179. DOI: https://doi.org/10.1145/3715275.3732148.

McCormack, Louise and Malika Bendechache (2025). "A comprehensive survey and classification of evaluation criteria for trustworthy artificial intelligence". In: *AI and Ethics* 5.3, pp. 1973–1994. DOI: https://doi.org/10.1007/s43681-024-00590-8.

McCormack, Louise, Malika Bendechache, et al. (2025). "Trust and transparency in AI: industry voices on data, ethics, and compliance". In: *AI & SOCIETY*, pp. 1–29. DOI: https://doi.org/10.1007/s00146-025-02654-7.

Montreal AI Ethics Institute (2025). *From Case Law to Code: Evaluating AI's Role in the Justice System*. Published online. URL: https://montrealethics.ai/from-case-law-to-code-evaluating-ais-role-in-the-justice-system/.

Morrison, Jacob Daniel et al. (2025). "Holistically Evaluating the Environmental Impact of Creating Language Models". In: *ArXiv* abs/2503.05804. DOI: https://doi.org/10.48550/arXiv.2503.05804. URL: https://api.semanticscholar.org/CorpusID:276902612.

Neil, Roland and Michael Zanger-Tishler (2025). "Algorithmic bias in criminal risk assessment: The consequences of racial differences in arrest as a measure of crime". In: *Annual Review of Criminology* 8. DOI: https://doi.org/10.1146/annurev-criminol-022422-125019.

OECD (2025). *AI in Justice Administration and Access to Justice: Governing with Artificial Intelligence*. Tech. rep. Paris: Organisation for Economic Co-operation and Development (OECD). URL: https://www.oecd.org/en/publications/2025/06/governing-with-artificial-intelligence_398fa287/full-report/ai-in-justice-administration-and-access-to-justice_f0cbe651.html.

Perera, Maneesha et al. (2025). "Indigenous peoples and artificial intelligence: A systematic review and future directions". In: *Big Data & Society* 12.2, p. 20539517251349170. DOI: https://doi.org/10.1177/20539517251349170.

Pham, Tri Minh Triet et al. (2025). "Time to Retrain? Detecting Concept Drifts in Machine Learning Systems". In: *2025 IEEE/ACM 47th International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*. IEEE, pp. 260–271. DOI: https://doi.org/10.1109/ICSE-SEIP66354.2025.00029.

Rief, Rachael M, Raven A Lewis, and D Michael Applegarth (2025). "In Pursuit of Fairness: A Research Note on Gender Responsivity and Racial Bias in Criminal

Justice Actuarial Risk Assessments". In: *Criminal justice policy review* 36.1-2, pp. 40–53. DOI: https://doi.org/10.1177/08874034241300162.

Sánchez de Ribera, Olga et al. (2025). "Prison violence: a latent class analysis of adult male offenders using the RisCanvi". In: *Psychology, Crime & Law*, pp. 1–19. DOI: https://doi.org/10.1080/1068316X.2025.2538147.

Silva, Alisson CC et al. (2025). "The analysis of criminal recidivism: a hierarchical model-based approach for the analysis of zero-inflated, spatially correlated recurring event data". In: *Journal of the Royal Statistical Society Series A: Statistics in Society*, qnaf061.

The Indegenous (2025). *Indigenous Representation in AI: 21st Century Implication*. Accessed: 2025-09-12. URL: https://www.theindegenous.org/indigenous-representation-in-ai-21st-century-implication.

Tong, Richard J et al. (2025). "A First-Principles Based Risk Assessment Framework and the IEEE P3396 Standard". In: *arXiv preprint arXiv:2504.00091*.

UK Justice (2025). *Artificial Intelligence in Our Justice System: Final Report*. Tech. rep. Justice UK. URL: https://files.justice.org.uk/wp-content/uploads/2025/01/29201845/AI-in-our-Justice-System-final-report.pdf.

Viljoen, Jodi L et al. (2025). "Are risk assessment tools more accurate than unstructured judgments in predicting violent, any, and sexual offending? A meta-analysis of direct comparison studies". In: *Behavioral sciences & the law* 43.1, pp. 75–113. DOI: https://doi.org/10.1002/bsl.2698.

Wang, Xiaoyang and Christopher C Yang (2025). "Enhancing Multi-Attribute Fairness in Healthcare Predictive Modeling". In: *arXiv preprint arXiv:2501.13219*.

Wenzelburger, Georg, Karen Yeung, and Kathrin Hartmann (2025). "Smart Justice? Making sense of the rise of algorithm-based pre-trial risk assessment in criminal justice through 'legal models'". In: *Digital Society* 4.2, p. 48. DOI: https://doi.org/10.1007/s44206-025-00194-7.

Yu, Guo et al. (2025). "Towards fairness-aware multi-objective optimization". In: *Complex & Intelligent Systems* 11.1, p. 50. DOI: https://doi.org/10.1007/s40747-024-01668-w.

Farayola, Michael Mayowa, Bendechache Malika, et al. (2026). "Beyond Calibration: Rethinking Algorithmic Fairness Through an Intersectional, Justice-Aware Lens". In.

Levin, Marc (n.d.). "881 NW 2d 749 (Wis. 2016). Judges rightly view sentencing as a weighty responsibility. They must consider not only the appropriate punishment for the offense but also the risk the offender poses, predicting the probability of the of-fender's recidivism. 2 A potential solution to this judicial anxiety has". In: ().

Xian, Ruicheng and Han Zhao (n.d.). "Efficient Post-Processing for Equal Opportunity in Fair Multi-Class Classification". In: ().