

FACULDADE DE CIÊNCIAS DA UNIVERSIDADE DO PORTO
FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Dynamic Morphing Augmentation for Algorithmic Fairness in Financial Models

Antónia Brito



Mestrado em Inteligência Artificial

Supervisors: Moisés Santos
Sérgio Jesus

Co-supervisor: Prof. Carlos Soares

July 31, 2026

Dynamic Morphing Augmentation for Algorithmic Fairness in Financial Models

Antónia Brito

Mestrado em Inteligência Artificial

Approved in oral examination by the committee:

President: Prof. Miriam Santos

Referee: Prof. Pedro Abreu

Referee: Prof. Moisés Santos

July 31, 2026

Resumo

Os modelos de classificação automatizados em domínios financeiros de elevado risco são cada vez mais implementados em ciclos de retroação, nos quais previsões passadas influenciam os dados de treino futuros, um fenómeno conhecido como degradação performativa da equidade. Nestas condições, as métricas de equidade padrão tornam-se pouco fiáveis: os modelos podem satisfazer a paridade demográfica através de atalhos patológicos, como o colapso para uma única classe, em que a exatidão agregada se mantém superficialmente aceitável enquanto as fronteiras de decisão colapsam por completo. As abordagens de auditoria tradicionais não conseguem expor estas falhas, deixando os sistemas de elevado risco vulneráveis a uma degradação não detetada. Para colmatar esta lacuna, esta tese introduz uma estrutura de dupla finalidade para auditar e mitigar o enviesamento de robustez demográfica em dados financeiros tabulares. O *FairMorph* é um método de aumento de dados dinâmico que resolve um problema discreto de Optimal Transport (OT) para emparelhar instâncias minoritárias com as suas contrapartes maioritárias geometricamente mais próximas, injetando amostras adversárias de travessia de fronteira que impõem invariância demográfica a instâncias estruturalmente semelhantes mas distintas em termos de grupo. O *TabMIST* é uma estrutura de auditoria geométrica que mede a deformação estrutural necessária para inverter as previsões ao longo de trajetórias de *morphing* entre classes, um sinal que, ao contrário das métricas agregadas, não pode ser manipulado por um colapso para a classe maioritária. Avaliado em três conjuntos de dados de fraude financeira sob condições estáticas e performativas, o *FairMorph* mantém um desempenho multiobjetivo estável onde os métodos de referência degradam: no conjunto de dados Feedzai Account Fraud, o seu Weighted Rank melhora de 2,17 para 1,83 sob stress performativo, enquanto um baseline Mixup DP colapsa de 2,81 para 5,25. As métricas padrão revelam-se pouco fiáveis sob retroação, com o rácio de Paridade Demográfica do baseline Vanilla a passar de +1,8% (estático) para -13,1% (performativo) no Folktables. O *TabMIST* expõe aquilo que estas métricas ocultam: no IBM Transaction Fraud, modelos com um rácio de Paridade Demográfica próximo de 1,000 revelam um $\text{TabMIST}_{\text{Avg}}$ a colapsar para perto de 0%, desmascarando um colapso para uma única classe em vez de uma equidade genuína, e auditorias ilustrativas revelam disparidades de robustez de até 9 pontos percentuais no *TabMIST* entre grupos demográficos de outra forma considerados equitativos. Esta abordagem combinada faz avançar a prática de Responsible Artificial Intelligence (RAI), ao permitir uma equidade algorítmica verificável a longo prazo em sistemas financeiros dinâmicos orientados por retroação.

Palavras-chave: Equidade Algorítmica, Performatividade Algorítmica, Morfismo de Conjuntos de Dados, Augmentação de Dados, Viés de Robustez, Detecção de Fraude Financeira

Abstract

Automated classification models in high-stakes financial domains are increasingly deployed in feedback loops where past predictions influence future training data, a phenomenon known as performative fairness degradation. Under these conditions, standard fairness metrics become unreliable: models can satisfy demographic parity through pathological shortcuts such as single-class default, where aggregate accuracy stays superficially acceptable while decision boundaries collapse entirely. Traditional auditing approaches fail to expose these failures, leaving high-stakes systems vulnerable to undetected degradation. To address this gap, this thesis introduces a dual-purpose framework for auditing and mitigating demographic robustness bias in tabular financial data. *FairMorph* is a dynamic augmentation method that solves a discrete Optimal Transport (OT) problem to pair minority instances with their geometrically closest majority counterparts, injecting adversarial boundary-crossing samples that enforce demographic invariance for structurally similar, group-distinct instances. *TabMIST* is a geometric auditing framework that measures the structural deformation required to flip predictions along inter-class morphing trajectories, a signal that, unlike aggregate metrics, cannot be gamed by majority-class collapse. Evaluated across three financial fraud datasets under static and performative conditions, *FairMorph* maintains stable multi-objective rank performance where baselines degrade: on the Feedzai Account Fraud dataset, its Weighted Rank improves from 2.17 to 1.83 under performative stress, while a Mixup DP baseline collapses from 2.81 to 5.25. Standard metrics prove unreliable under feedback, with the Vanilla baseline’s Demographic Parity ratio shifting from +1.8% (static) to −13.1% (performative) on Folktables. *TabMIST* exposes what these metrics conceal: on IBM Transaction Fraud, models with a Demographic Parity ratio near 1.000 show $\text{TabMIST}_{\text{Avg}}$ collapsing toward 0%, unmasking single-class default rather than genuine equity, and illustrative audits reveal robustness disparities of up to 9 percentage points in *TabMIST* between demographic groups otherwise judged equitable. This combined approach advances Responsible Artificial Intelligence (RAI) practice by enabling verifiable long-term algorithmic equity in dynamic, feedback-driven financial systems.

Keywords: Algorithmic Fairness, Algorithmic Performativity, Dataset Morphing, Data Augmentation, Robustness Bias, Financial Fraud Detection

UN Sustainable Development Goals

The United Nations Sustainable Development Goals (SDGs) provide a global framework for achieving a better and more sustainable future for all. It includes 17 goals¹ to address the world's most pressing challenges, including poverty, inequality, climate change, environmental degradation, peace, and justice. This thesis contributes to the advancement of responsible Artificial Intelligence in high-stakes financial systems, with a direct focus on algorithmic fairness, demographic equity, and trustworthy automated decision-making. These efforts align with the following Sustainable Development Goals:

The specific Sustainable Development Goals mentioned have the following names:

SDG 9 Build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation

SDG 10 Reduce inequality within and among countries

SDG 16 Promote peaceful and inclusive societies for sustainable development, provide access to justice for all and build effective, accountable and inclusive institutions at all levels

SDG	Target	Contribution	Performance Indicators and Metrics
9	9.5	FairMorph and TabMIST represent novel algorithmic research contributions that advance the technological capabilities of AI in the financial sector, directly supporting innovation in responsible machine learning.	Number of novel fairness-aware methods proposed; research and development expenditure as a proportion of GDP (9.5.1); researchers per million inhabitants (9.5.2).
	9.b	The framework supports domestic technology development by providing open, reproducible auditing and mitigation tools that can be adopted by financial institutions and policymakers to diversify and improve AI-driven decision systems.	Adoption rate of responsible AI tooling in financial institutions; proportion of medium and high-tech industry value added in total value added (9.b.1).

1. <https://sdgs.un.org/goals>

SDG	Target	Contribution	Performance Indicators and Metrics
10	10.2	By mitigating robustness bias that disproportionately penalises minority demographic groups in credit scoring and fraud detection, this work directly supports the social and economic inclusion of marginalised populations, irrespective of race, ethnicity, or economic status.	Reduction in group-level performance disparity metrics (e.g., $\text{TabMIST}_{\text{Avg}}$ across demographic groups); proportion of people living below 50% of median income disaggregated by group (10.2.1).
	10.3	FairMorph counteracts discriminatory patterns that can arise from historically biased training data, helping to eliminate algorithmically encoded discrimination in financial decision-making processes.	Fairness decay rate before and after FairMorph application; reduction in group accuracy disparity; proportion of population reporting discrimination (10.3.1).
	10.5	TabMIST provides interpretable sensitivity metrics that support the regulation and monitoring of automated financial models, contributing to the robustness and soundness of AI-driven financial systems.	$\text{TabMIST}_{\text{Avg}}$ and $\text{TabMIST}_{\text{Std}}$ stability scores across demographic groups; Financial Soundness Indicators (10.5.1).
16	16.3	By auditing and improving the fairness of automated classifiers in financial contexts, this framework supports equal access to financial services and just algorithmic outcomes for all demographic groups, promoting rule-of-law-consistent AI deployment.	Proportion of demographic groups achieving equitable classifier outcomes; switch Threshold disparity across groups as measured by TabMIST.
	16.6	The proposed auditing framework (TabMIST) promotes transparency and accountability in automated financial institutions, enabling stakeholders to detect and challenge discriminatory model behaviour, thereby fostering effective and accountable AI-driven institutions.	Model transparency score; percentage of financial institutions adopting algorithmic auditing practices; TabMIST as a measure of decision-boundary stability across groups.

Acknowledgements

To my supervisors, Moisés Santos, Sérgio Jesus and Carlos Soares, for guiding me throughout this project. Their experience, feedback, and encouragement allowed me to shape, what started as a chaotic literature review, into a full-fledged research project.

To Feedzai, namely Pedro Bizarro and Feedzai Research, for supporting and funding this research project through an internship that gave me the opportunity to work on a real-world problem with real-world stakes. To my manager, Iker Lopez, and colleagues at Feedzai Research, for welcoming me into the team, sharing their industry expertise, and making the experience as rewarding as it was formative.

To my parents, Isabel and J.Miguel, for their love and unwavering support; to my brother, Miguel, who has always annoyed me but looked out for me; to my grandparents, and their unconditional love and care.

To my four-legged furballs, Jimmy, Poppy, Teddy, Üglan, Rogerio, Constança, Mikinhas, Tonico and Xana, who always make my days better.

To all my old friends, and the new ones I made along the way. Thank you for always showing me new perspectives, supporting my creative work and having fun. I know I'll always be in good hands.

And to you, Pedro, for always cheering me up and pushing me to do my absolute best. I couldn't have done it without you.

Antónia Brito

Declaration of honour

I declare that this dissertation is my original work and has not previously been submitted or assessed in any other course or curricular unit, whether at this or any other institution. References to other authors (statements, ideas, thoughts), as well as the use of published works fully complies with the rules of attribution and are appropriately cited in the text and bibliography, in accordance with reference standards. I am aware that plagiarism and self-plagiarism constitute academic misconduct.

I further declare that I have not used Artificial Intelligence tools to produce any part(s) of this dissertation or, when such tools were used, their use and results are clearly indicated and were carefully reviewed for the detection of errors, inaccuracies, unfounded attributions, or other forms of content and argument bias, as well as for the determination of authorship.

I further declare that I have included studies carried out with multiple co-authors, having specified, with transparency and rigor, the contribution of each co-author; provided information on the possible existence of any other dissertation in which the same article has been included; reproduced the work in full with the permission of the journal in which it was published; and obtained the explicit written consent of all co-authors for the inclusion of the article.

Antónia Brito



“But how awful would that be? How terrible to live surrounded by the stark, sharp, hollowness of things that simply were enough?”

Auri, The Slow Regard of Silent Things

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Ethical Mandates of Responsible AI	2
1.3	Financial Fraud Context	3
1.4	Research Objectives	3
1.5	Thesis Outline	4
2	Literature Review	5
2.1	Algorithmic Fairness in Machine Learning	6
2.1.1	Group Fairness Metrics	6
2.1.2	Individual Fairness Metrics	8
2.1.3	Limitations and Mathematical Incompatibilities	8
2.2	Model Robustness in Machine Learning	10
2.2.1	Adversarial Robustness and Decision Boundary Probing	10
2.2.2	Margin-Based Robustness Frameworks	11
2.2.3	Limitations and Optimization Trade-offs	12
2.3	Intersection Between Fairness and Robustness	14
2.4	Evolution of Continuous Generative Modeling	18
2.4.1	Flow Matching: A Simulation-Free Paradigm	20
2.4.2	Conditional Flow Matching Formalization	20
2.4.3	Optimal Transport and Path Straightening	21
2.5	Final Remarks	22
3	FairMorph	23
3.1	Problem Scope	23
3.2	Dataset Morphing via Optimal Transport	25
3.3	Dynamic Augmented Training for Fair Robustness	26
3.4	FairMorph: Adversarial Augmented Training	29
4	TabMIST	33
4.1	Methodological Foundations	33
4.2	TabMIST: Morphing-based Sensitivity Metric	34
4.3	Evaluation of Demographic Robustness Bias	36
5	Experimental Setup	38
5.1	Datasets	38
5.1.1	Folktables Dataset	39
5.1.2	Feedzai Bank Account Fraud Dataset	39

5.1.3	IBM Benchmark Dataset	40
5.1.4	Robust Normalization	41
5.2	Model Architecture	41
5.3	Evaluation Metrics	42
5.4	Experimental ML Pipeline	42
6	Results and Discussion	46
6.1	RQ1: Do fairness metrics decay over performative loops in tabular distributions?	46
6.1.1	Label Feedback Loop Decay	47
6.1.2	Decay Over Time	48
6.2	RQ2: How does the proposed <i>FairMorph</i> framework compare against existing state-of-the-art mitigation strategies under performative stress?	48
6.3	RQ3: Does the introduction of the <i>TabMIST</i> framework provide tangible benefits to empirical model auditing?	50
6.4	Limitations	52
7	Conclusions and Future Work	54
7.1	Definitive Synthesis of Research Questions	54
7.2	Summary of Primary Contributions	55
7.3	Future Work	56
7.4	Concluding Remarks	56
	References	58
A	Supplemental Material	65
A.1	Multi-Objective Ranking Formalization	65
A.2	Directional Categories and Evaluation Framework	66
B	Tables of Results	67

List of Figures

2.1	Binary classification confusion matrix architecture.	6
3.1	Dataset generation process via linear optimal transport mapping on the Folktable Dataset. The transformation shifts the source dataset (purple) toward the target dataset (cyan), with stars marking respective centroids and arrows indicating the direction of optimal transport mapping.	26
3.2	Dynamic data augmentation pipeline, showcasing the generation of new synthetic data derived from Tabular Morpher interpolation between the vulnerable source and robust target datasets.	27
3.3	<i>FairMorph</i> data augmentation pipeline. At each epoch, the <i>Tabular Morpher</i> pairs source and target instances via Optimal Transport (OT) and generates intermediate synthetic snapshots along their linear interpolation trajectory. These snapshots are passed through the intermediate classifier to identify the adversarial samples, where the prediction first flips to the target label.	30
4.1	PCA projection of the data manifold highlighting the geometric positioning of stubborn (red circles) and weak (black circles) samples relative to the source and target clusters.	36
4.2	<i>Switch Threshold</i> distributions computed by <i>TabMIST</i> across demographic groups on the Folktables dataset, evaluated on a sample of 5,000 instances using an Multi-Layer Perceptron (MLP) model. Group A (left) represents the minority non-white demographic cohort and exhibits a pronounced skew toward higher threshold values. Group B (right) represents the majority white demographic cohort and displays a comparatively centered distribution.	37
6.1	Trade-off between predictive utility (F_1 -score) and Demographic Parity (DP). Each point represents the performance of a specific model configuration. Darker, solid circles denote performance in the static baseline environment, while lighter, hollow circles denote performance under the performative feedback loop. Dashed lines (arrows) indicate the directional decay of model performance over the performative horizon.	47
6.2	Multi-objective Weighted Rank stability across environments. The subplots trace the evolution of the Weighted Rank from the static baseline to the dynamic performative environment across the Folktable, Account Fraud, and IBM Transaction Fraud datasets.	49

- 6.3 Trade-off between predictive utility (F_1 -score) and fairness for IBM Transaction Fraud Data. Each point represents the performance of a specific model configuration. Darker, solid circles denote performance in the static baseline environment, while lighter, hollow circles denote performance under the performative feedback loop. Dashed lines (arrows) indicate the directional decay of model performance over the performative horizon. 51

List of Tables

2.1	Taxonomy of Mathematical Fairness Formalizations.	9
2.2	Taxonomy of Model Robustness Metrics and Boundary Probing.	13
2.3	Paradigms at the Intersection of Fairness and Robustness.	19
3.1	Comparison between sampling approaches and <i>FairMorph</i> augmentation averaged across multiple training phases of the Folktables dataset with the original ground truth. Values in bold highlight the best-performing method regarding accuracy, F1-score, and the ratio of fairness metrics across groups.	31
3.2	Hybrid Metric Rankings. Ranked over the average of the fairness ratios and performance metrics for the different methods.	32
6.1	Computational overhead of the <i>FairMorph</i> augmentation step across datasets, measured on a fixed sample size of 5,000 instances per training step. Morphing time is reported as the wall-clock time spent exclusively within the <code>fit_morph_linear</code> call. Overhead ratio is computed as the ratio of total augmented training time to total vanilla training time.	53
A.1	Weighting schema applied to the consolidated metric ranking matrix.	66
B.1	Folktable Dataset Results: Comparison between different bias mitigation approaches and FairMorph augmentation. Results averaged across training steps, where original ground truth was used (BASELINE). Values in bold highlight the best-performing method regarding accuracy, F1-score, and the ratio of fairness metrics across groups.	67
B.2	Folktable Dataset Results: Comparison between different bias mitigation approaches and FairMorph augmentation. Results averaged across training steps, where model predictions were used as labels (PERFORMATIVE). Values in bold highlight the best-performing method regarding accuracy, F1-score, and the ratio of fairness metrics across groups.	68
B.3	Account Fraud Dataset Results: Comparison between different bias mitigation approaches and FairMorph augmentation. Results averaged across training steps, where original ground truth was used (BASELINE). Values in bold highlight the best-performing method regarding accuracy, F1-score, and the ratio of fairness metrics across groups.	69
B.4	Account Fraud Dataset Results: Comparison between different bias mitigation approaches and FairMorph augmentation. Results averaged across training steps, where model predictions were used as labels (PERFORMATIVE). Values in bold highlight the best-performing method regarding accuracy, F1-score, and the ratio of fairness metrics across groups.	70

B.5	IBM Transaction Fraud Dataset Results: Comparison between different bias mitigation approaches and FairMorph augmentation. Results averaged across training steps, where original ground truth was used (BASELINE). Values in bold highlight the best-performing method regarding accuracy, F1-score, and the ratio of fairness metrics across groups.	71
B.6	IBM Transaction Fraud Dataset Results: Comparison between different bias mitigation approaches and FairMorph augmentation. Results averaged across training steps, where model predictions were used as labels (PERFORMATIVE). Values in bold highlight the best-performing method regarding accuracy, F1-score, and the ratio of fairness metrics across groups.	72
B.7	Folktable Dataset Metric Rankings. Ranked over the average of the fairness ratios and performance metrics for the different methods. (BASELINE)	72
B.8	Folktable Dataset Metric Rankings. Ranked over the average of the fairness ratios and performance metrics for the different methods. (PERFORMATIVE)	72
B.9	Account Fraud Dataset Metric Rankings. Ranked over the average of the fairness ratios and performance metrics for the different methods. (BASELINE)	73
B.10	Account Fraud Dataset Metric Rankings. Ranked over the average of the fairness ratios and performance metrics for the different methods. (PERFORMATIVE)	73
B.11	IBM Transaction Fraud Dataset Metric Rankings. Ranked over the average of the fairness ratios and performance metrics for the different methods. (BASELINE)	73
B.12	IBM Transaction Fraud Dataset Metric Rankings. Ranked over the average of the fairness ratios and performance metrics for the different methods. FaiR- $N_{0.5,0.5}$ and Mixup DP methods are not included since the model only predicts one class, therefore being useless. (PERFORMATIVE)	73
B.13	Folktable Dataset metric decay over time: Difference in results from the last training step and the first training step, where original ground truth was used (BASELINE). Values in bold highlight the lowest decay across methods regarding accuracy, F1-score, and the ratio of fairness metrics across groups.	74
B.14	Folktable Dataset metric decay over time: Difference in results from the last training step and the first training step, where model predictions were used as labels (PERFORMATIVE). Values in bold highlight the lowest decay across methods regarding accuracy, F1-score, and the ratio of fairness metrics across groups.	74
B.15	Account Fraud Dataset metric decay over time: Difference in results from the last training step and the first training step, where original ground truth was used (BASELINE). Values in bold highlight the lowest decay across methods regarding accuracy, F1-score, and the ratio of fairness metrics across groups.	74
B.16	Account Fraud Dataset metric decay over time: Difference in results from the last training step and the first training step, where model predictions were used as labels (PERFORMATIVE). Values in bold highlight the lowest decay across methods regarding accuracy, F1-score, and the ratio of fairness metrics across groups.	75
B.17	IBM Transaction Fraud Dataset metric decay over time: Difference in results from the last training step and the first training step, where original ground truth was used (BASELINE). Values in bold highlight the lowest decay across methods regarding accuracy, F1-score, and the ratio of fairness metrics across groups.	75

B.18 IBM Transaction Fraud Dataset metric decay over time: Difference in results from the last training step and the first training step, where model predictions were used as labels (PERFORMATIVE). Values in bold highlight the lowest decay across methods regarding accuracy, F1-score, and the ratio of fairness metrics across groups.	75
---	----

List of Acronyms

AI	Artificial Intelligence
CF	Counterfactual Fairness
CFM	Conditional Flow Matching
CNF	Continuous Normalizing Flow
DCFDG	Disentanglement for Counterfactual Fairness-aware Domain Generalization
DNN	Deep Neural Network
DP	Demographic Parity
DR	Distributionally Robust
DTW	Dynamic Time Warping
EO	Equalized Odds
FM	Flow Matching
FN	False Negatives
FP	False Positives
FTA	Fairness Through Awareness
GANs	Generative Adversarial Networks
GDPR	General Data Protection Regulation
I-CFM	Independent Conditional Flow Matching
ML	Machine Learning
MLP	Multi-Layer Perceptron
NN	Neural Network
ODE	Ordinary Differential Equation
OT	Optimal Transport
OT-CFM	Optimal Transport Conditional Flow Matching
PCA	Principal Component Analysis
RAI	Responsible Artificial Intelligence
RB	Robustness Bias
RF	Recourse Fairness
RFR	Robust Fairness Regularization
RGG	Robustness Generalization Gap
SLSD	Shared Latent Space-based Debiasing

TN	True Negatives
TP	True Positives

Chapter 1

Introduction

This master’s research is grounded in the principles of Responsible Artificial Intelligence (**RAI**), to ensure theoretical and methodological rigor when applying Artificial Intelligence (**AI**) to sensitive domains. This chapter contextualizes the importance of robustness, fairness, and their interdependence in real-world deployment. Additionally, it outlines the primary objectives of this work.

1.1 Motivation

AI technologies are achieving remarkable advances in critical sectors, including medical imaging (Ahmad, Eckert, and Teredesai 2018), autonomous driving (Yurtsever et al. 2020), criminal justice (Završnik 2021), financial services (Abdallah, Maarof, and Zainal 2016), and hiring (Wilson et al. 2021). This proliferation has brought substantial improvements in efficiency and data processing capabilities. However, the rapid adoption of these systems introduces growing responsibilities grounded in societal and ethical challenges. High-profile cases have revealed how brittle **AI** systems can perpetuate and even amplify existing biases, leading to systematic discrimination despite no intent to cause harm (Obermeyer et al. 2019; Koenecke et al. 2020).

In response, the field of **RAI** has emerged as a focal discipline to ensure **AI** systems align with fundamental human values (Dignum 2019). **RAI** represents an organic evolution of the field, moving beyond purely technical optimization to embrace the integration of ethical principles throughout the **AI** life cycle. This paradigm shift is now being codified into applicable actions by global awareness and a regulatory wave. Despite these regulatory mandates, a significant gap remains in the methodological unification of **RAI** principles. Current literature predominantly treats properties such as fairness and robustness in isolation. Fairness research focuses on mitigating historical disparities, while robustness research targets stability against adversarial attacks or distribution shifts. Recent findings, however, suggest that these concepts should be integrated, as specific groups can be disproportionately affected by robustness issues a phenomenon known as Robustness Bias (**RB**) (Nanda et al. 2021; Sharma et al. 2021). This thesis will address the limitations of both siloed and novel approaches, drawing from concepts of Optimal Transport (**OT**) and dataset morphing to better reconcile the trade-offs between equal treatment and stability.

1.2 Ethical Mandates of Responsible AI

While the **AI** general adoption motivates this practice, this emerging legislative framework defines the boundaries of acceptable operation. Consequently, ethical concepts like fairness and robustness are no longer being regarded as optional; they are concrete business and engineering requirements.

The enforcement of the General Data Protection Regulation (**GDPR**) (**European Parliament and Council of the European Union 2016**) in 2018 established an initial baseline, creating a legal right for individuals to demand explanations for automated decisions. It additionally defined sensitive data, which includes racial and ethnic origin, religious beliefs, and sexual orientation. More recently, the 2024 European **AI** Act has formalized a risk-based approach, explicitly listing "robustness, accuracy, and cybersecurity" as prerequisites for high-risk applications. **RAI** is therefore the engineering response to this legislative mandate, translating abstract legal requirements into more verifiable system properties.

Fairness Legally, fairness refers to the principle of non-discrimination, ensuring that **AI** decisions do not create inequitable outcomes for protected groups. While European legislation explicitly prohibits discrimination based on the sensitive attributes described above, achieving this mathematically is complex. As detailed in Chapter 2, merely removing these attributes is insufficient due to proxy variables that indirectly encode historical bias. **AI** systems can unknowingly perpetuate these biases (O'Connor and Liu 2024), as seen in the COMPAS recidivism algorithm (Dressel and Farid 2018).

Notably, Chouldechova (2017) subsequent analysis of COMPAS revealed that no single fairness metric can simultaneously satisfy multiple demographic parity constraints without incurring unavoidable trade-offs, underscoring that fairness is not merely a technical fix but an inherently value-laden design choice requiring explicit specification and rigorous auditing. This impossibility transforms fairness from a technical problem into an irreducibly normative one, where no single metric can simultaneously guarantee non-discrimination across all dimensions. Thus, the ethical mandate for fairness demands mathematically rigorous metrics to audit systems against the systematic discrimination prohibited by law, alongside transparent acknowledgment of the unavoidable tensions embedded within competing fairness objectives.

Robustness Robustness is traditionally viewed as a security metric, a system's ability to withstand adversarial attacks (Chakraborty et al. 2021) or distribution shifts (Taori et al. 2020). In the context of **RAI**, robustness is a necessity. A system that is "brittle" is inherently untrustworthy, meaning it is fragile in the face of adversarial perturbations or distribution shifts, losing predictive performance under environmental deviation. If a model's ethical properties and performance degrade under slight perturbations, it cannot be deemed safe for deployment. Crucially, this thesis views robustness through an intersectional lens: if an **AI** system is less robust for minority groups, it also constitutes a violation of the fundamental principle of fairness.

While the regulatory mandate is clear, systems must be both fair and robust; satisfying these constraints simultaneously remains an open research challenge. As the following chapters will demonstrate, naive attempts to secure robustness do not necessarily guarantee fairness between the different groups in the data, and rigid fairness constraints can destabilize model performance. This thesis aims to bridge this gap, proposing evaluation practices and a methodology that satisfy these dual ethical mandates through dynamic augmentation.

1.3 Financial Fraud Context

To validate the theoretical contributions of this thesis, the research focuses on the domain of financial fraud detection in collaboration with Feedzai. This sector serves as a high-stakes automated risk assessment application, providing a critical testbed for analyzing the interactions between fairness and robustness.

Financial access is a fundamental enabler of financial inclusion. In this context, the cost of a False Positives (FP) (incorrectly flagging a legitimate transaction as fraud) is not merely a statistical error but a denial of service that can be detrimental to an individual's well-being. A fairness-aware system must therefore account for this asymmetry, ensuring that protective measures do not disproportionately burden vulnerable populations or threaten access to essential services.

The financial environment is inherently dynamic and adversarial, presenting an increased challenge for points of failure in predictive models:

- **Adversarial Evolution:** Fraud detection involves an active adversary. Fraudsters continuously probe the decision boundary, evolving their tactics to evade detection, which constitutes a real-world adversarial attack.
- **Distribution Shifts:** The underlying data distribution is in constant flux due to economic factors such as inflation, seasonal spending behaviors, and evolving payment trends.
- **Performative Prediction:** As defined by Perdomo et al. (2020), financial models exhibit performativity, where the system's decisions alter the future data distribution. This phenomenon may cause a concept shift, where the underlying relationships between the model's features and outcomes change over time (Makowski 2024; Mishler and Dalmaso 2022). This feedback loop can potentially cement biases against specific demographic groups or other negative patterns.

This combination of high impact and environmental volatility makes financial fraud detection a good candidate for studying the intersection of algorithmic fairness and model robustness.

1.4 Research Objectives

This thesis proposes a framework to bridge the gap between algorithmic fairness and model robustness through two primary objectives:

Objective 1: Characterize and Audit Robustness Bias in Dynamic Environments: To quantify how performative distribution shifts disproportionately impact protected subgroups in non-stationary financial systems, thereby establishing a baseline robustness gap. By introducing *TabMIST* as a geometric auditing instrument to probe the micro-level topography of classifier decision boundaries across demographic groups.

Objective 2: Develop and Validate Boundary-Targeted Morphing Mitigation: To design an adaptive data augmentation framework (*FairMorph*) leveraging optimal transport and boundary-crossing trajectories. This objective evaluates whether targeted boundary reinforcement can navigate the trade-off between stability and group fairness while equalizing the burden of recourse across demographic cohorts.

To systematically fulfill these objectives, this thesis addresses four central research questions:

- **RQ1: Do fairness metrics decay over performative loops in tabular distributions?**

Investigates whether group-level equity criteria satisfied during static training persist when deployed in active feedback loops where model predictions alter subsequent data generation.

- **RQ2: How does the proposed FairMorph framework compare against existing state-of-the-art mitigation strategies under performative stress?**

Evaluates the multi-objective rank stability and decision-boundary resilience of *FairMorph* relative to in-processing regularizers and manifold data-smoothing baselines, with particular attention to behavior under long-term feedback.

- **RQ3: Does the introduction of the TabMIST framework provide tangible benefits to empirical model auditing?**

Assesses whether probing decision boundaries via continuous, sample-level dataset morphing uncovers topological anomalies that aggregate, macro-level metrics fail to detect.

1.5 Thesis Outline

The remainder of this thesis systematically addresses the research objectives across six subsequent chapters. Chapter 2 establishes the theoretical foundations of algorithmic fairness, structural robustness, and OT-based continuous flow paradigms. Chapter 3 details the mechanics of the proposed *FairMorph* dataset morphing framework for boundary reinforcement, complemented by Chapter 4, which introduces the *TabMIST* diagnostic suite to audit decision surfaces. Chapter 5 outlines the experimental pipeline, baseline configurations, and tabular datasets, which form the foundation for the empirical performance evaluations presented in Chapter 6. Finally, Chapter 7 summarizes the primary contributions to RAI and outlines prospective avenues for future research.

Chapter 2

Literature Review

This chapter establishes the theoretical foundations necessary to analyze the intersection of algorithmic fairness and model robustness. We review the state-of-the-art in three core areas: algorithmic fairness metrics, model robustness against distribution shifts and performative loops, and the evolution of continuous generative modeling via Flow Matching (FM) and OT. By studying the limitations of current methodologies in these fields, this thesis identifies the specific gaps that arise when these properties interact, providing the necessary mathematical and conceptual runway for our proposed framework.

The remainder of this chapter is organized as follows:

- **Section 2.1: Algorithmic Fairness** reviews the formal mathematical taxonomies of group, individual, and recourse-based fairness metrics, while detailing the fundamental theorems that govern their statistical incompatibilities.
- **Section 2.2: Model Robustness** examines classifier stability under adversarial interference and domain drift, analyzing the geometric properties of margin distributions and the structural mechanics driving the robust generalization gap.
- **Section 2.3: The Intersection of Fairness and Robustness** isolates the critical points of systemic tension, trade-offs, and open vulnerabilities that materialize when these two pillars are optimized concurrently in dynamic or data-poisoned environments.
- **Section 2.4: Evolution of Continuous Generative Modeling** traces the mathematical transition from continuous normalizing flows to simulation-free Conditional Flow Matching (CFM) and Optimal Transport Conditional Flow Matching (OT-CFM), establishing the geometric foundations that inspire our core methodology.
- **Section 2.5: Final Remarks** synthesizes the findings in this literature review and introduces the motivation for our main contributions.

2.1 Algorithmic Fairness in Machine Learning

Algorithmic fairness seeks to identify, quantify, and mitigate systemic bias in Machine Learning (ML) models by enforcing formal mathematical constraints on their optimization objectives and predictions. While specific definitions vary drastically based on underlying normative, legal, and ethical paradigms (*e.g.*, focusing on equality of outcome versus equality of treatment), contemporary fairness research generally partitions these constraints into two primary analytical families: Group Fairness and Individual Fairness (Mehrabi et al. 2021).

2.1.1 Group Fairness Metrics

Group fairness metrics evaluate algorithmic equity by requiring that specific statistical criteria of a classifier remain balanced across distinct population subgroups defined by a protected or sensitive demographic attribute $a \in \{0, 1\}$ (*e.g.*, race, gender, or age).

		Ground Truth	
		P	N
Prediction	$\hat{P}$	True Positive	False Positive
	$\hat{N}$	False Negative	True Negative

Figure 2.1: Binary classification confusion matrix architecture.

To establish a rigorous mathematical foundation for these group-level evaluations, it is necessary to ground them within the classical framework of binary classification. Let $Y \in \{0, 1\}$ represent the empirical ground-truth label, and let $\hat{Y} \in \{0, 1\}$ denote the binary prediction output by a classifier model $f(x)$. The intersection of these target vectors yields the standard confusion matrix layouts illustrated in Figure 2.1, defining True Positives (TP), FP, True Negatives (TN), and False Negatives (FN). Group fairness frameworks track how these base error distributions fluctuate when conditioned on the sensitive attribute a .

Independence (Demographic Parity) The independence criterion (Calders, Kamiran, and Pechenizkiy 2009), widely formalized as Demographic Parity (DP) or Statistical Parity, demands that the classifier’s ultimate prediction $\hat{Y}$ be statistically independent of the sensitive attribute a . This constraint implies that the marginal probability of receiving a positive classification must remain invariant across all protected demographic cohorts:

$$P(\hat{Y} = 1 \mid a = 0) = P(\hat{Y} = 1 \mid a = 1) = P(\hat{Y} = 1). \quad (2.1)$$

In an operational context such as automated hiring, where $\hat{Y} = 1$ denotes a favorable employment recommendation, satisfying **DP** dictates that the selection rate must be strictly equalized between groups. For instance, if a model yields a positive selection rate of 20% for the privileged group ($a = 1$), it must yield an identical 20% selection rate for the unprivileged group ($a = 0$), completely abstracting away from the underlying distribution of true qualifications across those groups.

Separation (Equalized Odds & Equal Opportunity) The separation criterion addresses the qualification disparities ignored by independence by requiring that the classifier's prediction $\hat{Y}$ be conditionally independent of the sensitive attribute a , given the true ground-truth label Y :

$$P(\hat{Y} = 1 \mid Y = y, a = 0) = P(\hat{Y} = 1 \mid Y = y, a = 1), \quad \forall y \in \{0, 1\}. \quad (2.2)$$

Operationally, separation enforces the simultaneous balance of **TP** and **FP** rates across all demographic cohorts, a condition known as Equalized Odds (**EO**) (Hardt, Price, and Srebro 2016). A widely utilized relaxation of this metric is Equal Opportunity, which isolates and enforces this conditional independence constraint strictly for the positive class ($Y = 1$):

$$P(\hat{Y} = 1 \mid Y = 1, a = 0) = P(\hat{Y} = 1 \mid Y = 1, a = 1). \quad (2.3)$$

In a consumer lending scenario, Equal Opportunity dictates that creditworthy individuals ($Y = 1$) maintain an equal probability of loan approval regardless of their demographic classification, thereby protecting qualified minority applicants from systemic exclusion.

Sufficiency (Predictive Parity) Sufficiency turns the conditional directionality around, requiring the conditional independence of the true label Y and the sensitive attribute a , conditioned directly on the model's prediction score $\hat{Y}$:

$$P(Y = 1 \mid \hat{Y} = 1, a = 0) = P(Y = 1 \mid \hat{Y} = 1, a = 1). \quad (2.4)$$

This constraint mandates that the precision (or positive predictive value) of the system remains uniform across all subgroups. When sufficiency is satisfied, a positive classification carries an identical statistical meaning and likelihood of actual validation across both privileged and unprivileged groups, a concept intrinsically bound to predictive model calibration.

Burden (Recourse Fairness (RF)) Shifting away from the purely rate-based statistical metrics detailed above, contemporary fairness literature has introduced the concept of the *Burden of Recourse*. As formalized by Sharma et al. (2021), burden quantifies the minimum geometric effort, measured as the distance from an input coordinate to the active decision boundary B , that an individual assigned a negative prediction must exert to mutate their features and achieve a positive classification outcome. A classifier satisfies **RF** if the expected geometric burden for rejected

individuals is balanced across sensitive groups:

$$\mathbb{E}[d(x, \mathcal{B}) \mid \hat{Y} = 0, a = 0] = \mathbb{E}[d(x, \mathcal{B}) \mid \hat{Y} = 0, a = 1], \quad (2.5)$$

where $d(x, \mathcal{B})$ represents the shortest distance from input feature vector x to the decision manifold $\mathcal{B}$. This metric guarantees that even if base denial rates are statistically equalized, one subgroup does not confront structurally higher geometric or operational barriers to reversing an adverse decision.

2.1.2 Individual Fairness Metrics

In contrast to group-level abstractions that aggregate statistics over entire populations, individual fairness frameworks focus on ensuring consistent and equitable treatment at the level of individual instances.

Fairness Through Awareness (FTA) Dwork et al. (2012) formalized the foundational individual fairness axiom that “similar individuals should be treated similarly.” This paradigm requires the explicit definition of a task-specific similarity metric $d(\cdot, \cdot)$ operating in the input feature space, alongside a distance metric $D(\cdot, \cdot)$ tracking variations in the output predictive distribution space. A mapping mechanism M satisfies individual fairness if and only if:

$$D(M(x), M(x')) \leq d(x, x'), \quad (2.6)$$

for all individual pairs x, x' . While theoretically elegant, the primary obstacle in deploying FTA resides in the difficulty of constructing a globally objective similarity metric d that accurately isolates legitimate qualifications without absorbing or reinforcing historical human biases.

Counterfactual Fairness (CF) Proposed by Kusner et al. (2017), a predictor $\hat{Y}$ satisfies CF if, for any individual instance characterized by observed features $X = x$ and native sensitive attribute a , the resulting prediction remains completely invariant had the individual belonged to an alternative demographic group a' , adjusting for all intermediate causal pathways:

$$P(\hat{Y}_{a \leftarrow a} = y \mid X = x, a = a) = P(\hat{Y}_{a \leftarrow a'} = y \mid X = x, a = a). \quad (2.7)$$

By relying on causal graph structures, CF distinguishes between discriminatory data-generating pathways and justifiable causal mechanisms, capturing subtle contextual biases that purely correlational, group-level metrics cannot fully isolate.

2.1.3 Limitations and Mathematical Incompatibilities

The proliferation of diverse mathematical definitions underscores that it is mathematically impossible to construct a model that satisfies all notions of equity simultaneously.

Table 2.1: Taxonomy of Mathematical Fairness Formalizations.

Fairness Paradigm	Core Metric	Mathematical Definition	Primary Objective
Group Fairness	Independence (DP)	$\hat{Y} \perp a$	Equalize selection rates
	Separation (EO)	$\hat{Y} \perp a \mid Y$	Equalize error rates (TP/FP)
	Sufficiency	$Y \perp a \mid \hat{Y}$	Equalize predictive precision
Recourse Fairness	Boundary Burden	$\mathbb{E}[d(x, B) \mid \hat{Y} = 0, a]$	Equalize effort required for recourse
Individual Fairness	Awareness (FTA)	$D(M(x), M(x')) \leq d(x, x')$	Consistent mapping of similar profiles
	Counterfactual (CF)	$\hat{Y}_{a \leftarrow a} = \hat{Y}_{a \leftarrow a'}$	Invariance under counterfactual switch

The Impossibility Theorems The foundational works of Chouldechova (2017) and Kleinberg, Mullainathan, and Raghavan (2016) established that except in scenarios where base outcome rates are perfectly identical across groups ($P(Y = 1 \mid a = 0) = P(Y = 1 \mid a = 1)$), it is mathematically impossible to simultaneously satisfy Independence (**DP**), Separation (**EO**), and Sufficiency (Predictive Parity). For example, if the baseline qualification rate varies between demographic groups, an algorithmic system that maintains predictive sufficiency must necessarily generate disparate error profiles across those cohorts, directly violating the separation criterion. This mathematical lock forces stakeholders to make explicit, context-dependent decisions based on the specific ethical and operational constraints of the deployment domain.

Incompatibility of Analytical Perspectives Beyond structural statistical conflicts, different mathematical proxies protect fundamentally competing interests. As emphasized by Sharma et al. (2021), a classifier can be fully optimized to satisfy classical group error constraints (such as **EO**) while completely failing to provide an equitable burden of recourse. For instance, even if error rates are balanced, the decision manifold can be oriented such that rejected minority individuals end up positioned significantly deeper within the negative region than rejected majority individuals. Consequently, the minority group faces structurally higher barriers to reverse adverse outcomes.

These persistent contradictions prove that algorithmic fairness cannot be treated as a pure, context-free optimization task. Mathematical abstractions like separation or independence capture only isolated fragments of the complex social, legal, and historical landscape of equity. Because these metrics are anchored in competing ethical principles, the selection of a fairness regularizer is fundamentally a normative decision rather than a technical one. Practitioners must explicitly align their choice of mathematical criteria with the broader socio-legal goals of the domain. To aid this selection process, frameworks such as the one proposed by Saleiro et al. (2018) organize these competing definitions into a programmatic taxonomy, pairing metrics with distinct context-dependent policy goals to guide practitioners toward the most appropriate criteria.

2.2 Model Robustness in Machine Learning

While **ML** models frequently achieve exceptional predictive utility within highly controlled testing environments, their performance routinely degrades or collapses when deployed in real-world settings. Freiesleben and Grote (2023) formally define robustness as the “relative stability of a robustness target with respect to specific interventions on a modifier.” Evaluated through a geometric lens, robustness characterizes a model’s capacity to maintain reliable, invariant predictive representations despite exposure to semantic noise, out-of-distribution shifts, or targeted adversarial interference.

The necessity of modeling structural robustness is highlighted by several pervasive failure modes that manifest throughout real-world deployment horizons:

- **Shortcut Learning:** Deep models frequently exploit confounding empirical artifacts or spurious correlations present within the training manifold (*e.g.*, hospital-specific tokens or text tags on medical X-rays) rather than capturing true causal features, resulting in catastrophic generalization failures when evaluated in novel clinical environments (Zech et al. 2018).
- **Natural Distribution Shifts:** Classifiers optimized on pristine datasets often degrade abruptly when processing lower-quality real-world inputs, a vulnerability illustrated by clinical diabetic retinopathy screening models failing under poor or uncalibrated field lighting conditions (Beede et al. 2020).
- **Performative Feedback Loops:** Performative shifts materialize when an algorithmic prediction loop dynamically alters the underlying data-generating distribution over successive operational horizons. For example, when a credit risk model assigns higher interest rates to structurally marginalized groups, it actively exacerbates their default risk, establishing a self-fulfilling cycle of performance degradation and systemic bias reinforcement (Perdomo et al. 2020).
- **Adversarial Vulnerabilities:** High-dimensional decision boundaries are highly susceptible to minute, targeted perturbations that remain completely imperceptible to human observers but deceive neural network representations. For instance, by introducing a small, optimized color patch into a physical scene, the visual perception models of autonomous vehicles can be completely paralyzed (Rajan et al. 2019).

2.2.1 Adversarial Robustness and Decision Boundary Probing

The most prominent subfield of robustness research centers on adversarial examples, where input vectors are intentionally modified by worst-case noise configurations optimized to deceive Deep Neural Network (**DNN**) architectures (Goodfellow, Shlens, and Szegedy 2014). Unlike random isotropic noise, these perturbations are explicitly computed to exploit the high-dimensional curvature and fragile local margins of the model’s decision boundary.

To bridge the analytical gap separating abstract, unconstrained adversarial noise from semantically coherent input variations, the *tsMIST* framework (Brito et al. 2025) shifts the auditing paradigm from random pixel- or feature-level budget attacks toward structured, sample-level manifold morphing. Traditional adversarial frameworks typically deploy mathematical optimization loops to identify an arbitrary, minimum-distance vector capable of pushing an instance across a decision surface (Moosavi-Dezfooli, Fawzi, and Frossard 2016; Szegedy et al. 2013). In contrast, *tsMIST* systematically explores decision topography by constructing continuous linear trajectories bridging empirical inter-class data pairs.

The core execution of this morphing framework relies on isolating borderline candidate pairs, which are defined as instances from opposing ground-truth classes that share high structural or geometric proximity, and performing deterministic interpolation between them to map out the local decision surface. Along each calculated interpolation trajectory, the framework isolates the *Switch Threshold*, which pinpoints the normalized coordinate ($\alpha \in [0, 1]$) where the classifier’s predicted label flips from the native source class to the destination target class.

By aggregating these localized threshold transitions over a representative testing population, the framework derives two macro-level diagnostic metrics:

1. $tsMIST_{Avg}$: Measures the expected location of the decision boundary across the trajectory population. A score approaching 50% indicates a highly balanced, symmetric classification margin, whereas extreme values skewed toward 0% or 100% expose brittle, easily compromised margins.
2. $tsMIST_{Std}$: Quantifies the spatial variance of these boundary crossings, tracking consistency and uniformity across distinct sub-regions of the high-dimensional feature space.

2.2.2 Margin-Based Robustness Frameworks

To explicitly mitigate model sensitivity to out-of-distribution perturbations, extensive research focuses on optimizing classification margins (Elsayed et al. 2018; Nanda et al. 2021; Bartlett, Foster, and Telgarsky 2017). Geometrically, the margin represents the shortest distance separating an empirical data coordinate from the active decision boundary. Intuitively, data points that possess expanded classification margins exhibit higher structural robustness, requiring a larger perturbation vector to displace them across the decision threshold and into a misclassification zone.

First-Order Proxy Distances and Boundary Bounds While classification margins are easily calculated for linear learning paradigms via closed-form geometry (e.g., Support Vector Machines), computing the exact distance $d(x)$ to the decision boundary for DNNs is computationally intractable due to the highly non-convex, winding topology of deep decision surfaces. To bypass this computational bottleneck, Elsayed et al. (2018) introduced a first-order Taylor series approximation to estimate local boundary distance. Let the pairwise decision boundary separating class i and class j be formalized as $\mathcal{D}_{(i,j)} \triangleq \{x \mid f_i(x) = f_j(x)\}$. The estimated geometric distance of an

intermediate activation vector x^l at layer l to this decision surface is formulated as:

$$d_{f,(i,j)}(x^l) = \frac{f_i(x^l) - f_j(x^l)}{\|\nabla_{x^l} f_i(x^l) - \nabla_{x^l} f_j(x^l)\|_2}, \quad (2.8)$$

where $f_i(x^l)$ represents the unnormalized output logit calculated for class i .

Because exact boundary tracking remains fundamentally intractable at scale, alternative frameworks proposed by Nanda et al. (2021) focus on bounding the true boundary distance $d(x)$ within rigorous upper and lower limits. The *Upper Bound* is derived empirically by generating an optimized adversarial counterpart $\tilde{x}$ via iterative projection algorithms like *DeepFool* (Moosavi-Dezfooli, Fawzi, and Frossard 2016), asserting that the physical distance between the native sample and its counterfactual establishes a strict ceiling: $\|x - \tilde{x}\| \geq d(x)$. Conversely, the corresponding *Lower Bound* is constructed through the deployment of Randomized Smoothing techniques (Salman et al. 2019), which calculate a provable probabilistic radius δ_x within which the model’s prediction is mathematically guaranteed to remain invariant, establishing that $d(x) \geq \delta_x$.

From Minimum Margins to Margin Distributions Early statistical generalization bounds focused primarily on optimizing the minimum margin, which isolates the worst-case distance of the closest empirical coordinate to the classification boundary (Bartlett, Foster, and Telgarsky 2017). However, Lyu, Wang, and Zhou (2022) demonstrate that optimizing solely for the minimum margin ignores critical geometric information regarding the broader data manifold.

Instead, they argue that frameworks must prioritize optimizing the entire margin distribution. To achieve this, they introduce a convex loss function that enforces a dedicated “zero-loss band,” explicitly penalizing both underfitting (where local margins are structurally too narrow) and overfitting (where margins are pushed to excessively large, uninformative volumes).

The Robustness Generalization Gap (RGG) The necessity of tracking margin density is connected to the phenomenon of overfitting in overparameterized regimes. Zhang et al. (2016) demonstrated that deep architectures can achieve zero training error even when optimized on completely randomized labels, proving that empirical training loss is an insufficient proxy for true generalization. Y. Jiang et al. (2018) formalized this discrepancy as the **RGG**, the persistent mismatch observed between training set robustness and test set robustness under distribution shifts. To accurately predict and close this gap, they utilize the structural characteristics of margin distributions evaluated across multiple intermediate hidden layers of the network rather than evaluating the final output logits in isolation.

2.2.3 Limitations and Optimization Trade-offs

While margin maximization serves as a powerful regularization tool, it introduces significant operational trade-offs that complicate Neural Network (NN) optimization. Kim et al. (2023) demonstrate that maximizing classification margins beyond a critical threshold yields severe diminishing

Table 2.2: Taxonomy of Model Robustness Metrics and Boundary Probing.

Robustness Paradigm / Metric	Core Mathematical Formulation / Proxy	Primary Operational Objective	Diagnostic & Topographical Role
Switch Threshold (α)	$\alpha = \min_{1 \leq i \leq n} i \text{ s.t. } \hat{Y}_i \neq Y^{(\text{source})}$	Identify the exact interpolation coordinate where a prediction flips.	Measures localized edge-case sensitivity along an inter-class trajectory.
$tsMIST_{\text{Avg}}$	$\frac{1}{\delta} \sum_{k=1}^{\delta} \alpha_k$	Quantify the global orientation and stability of classification margins.	Scores near 50% indicate symmetric margins; skewed scores reveal boundary fragility/rigidity.
$tsMIST_{\text{Std}}$	$\sqrt{\frac{1}{\delta} \sum_{k=1}^{\delta} (\alpha_k - tsMIST_{\text{Avg}})^2}$	Measure the geometric consistency of the boundary across the data manifold.	Low variance indicates uniform robustness; high variance exposes region-dependent vulnerabilities.
First-Order Proxy Distance	$\frac{d_{f,(i,j)}(x^d)}{\frac{f_i(x^d) - f_j(x^d)}{\ \nabla_{x^d} f_i(x^d) - \nabla_{x^d} f_j(x^d)\ _2}} =$	Estimate the geometric distance to the decision boundary using first-order Taylor approximation.	Bypasses non-convex computational tractability limits in (DNNs).
Boundary Bounds	$\ x - \tilde{x}\ _{\text{adversarial}} \geq d(x) \geq \delta_{\text{smoothing}}$	Establish provable mathematical limits on local boundary distances.	Uses adversarial examples (Ceiling) and randomized smoothing radii (Floor) to sandwich the true margin.
Margin Distribution	Convex loss function enforcing an optimized “zero-loss band”	Maximize and smooth the density of margins across the entire data mass.	Rejects outlier-heavy minimum margins to prevent robust generalization gaps.

returns and actively damages the **RGG**, indicating the existence of an optimal structural “sweet spot” for margin regularizers.

Furthermore, Y. Jiang et al. (2018) criticize classic generalization bounds that rely exclusively on output layers (Bartlett, Foster, and Telgarsky 2017). They show that within deep architectures, margin distributions must maintain spatial consistency across hidden representations to guarantee true structural robustness under deployment shifts.

Finally, Lyu, Wang, and Zhou (2022) note that methods focused entirely on worst-case minimum margins overlook the mass of the underlying data distribution. This oversight can lead to highly suboptimal decision surfaces that protect isolated outliers while rendering the average operational case brittle and unstable.

2.3 Intersection Between Fairness and Robustness

As seen in the past literature, the concepts of fairness and robustness have been extensively studied separately, with fairness studies concentrating on mitigating historical bias and disparities in the data, while robustness research focuses on mitigating data noise, adversarial attacks, and distribution shift. Approaches aimed at the intersection of fairness and robustness can diverge into multiple research lines.

Sample Selection Bias and Distributional Shifts In real-world deployment horizons, training distributions frequently suffer from sample selection bias, where certain demographic cohorts are systematically underrepresented, or covariate shift, where the input feature topology undergoes significant mutation post-deployment. Consequently, a model engineered to satisfy rigid fairness constraints on a static training set often loses its capacity to maintain equitable predictions when evaluated on unseen testing manifolds (Du and Wu 2021; Mandal et al. 2020; Rezaei et al. 2020). To resolve selection bias under unknown selection probabilities, Du and Wu (2021) introduce a framework that couples reweighting estimation with a min-max Distributionally Robust (DR) optimization strategy. This approach bounds prediction errors by minimizing the worst-case loss over a structured Wasserstein ball (Blanchet, Kang, and Murthy 2019) constructed around suspected selection probabilities. Similarly, Mandal et al. (2020) formulate a min-max multi-objective function designed to minimize a DR loss while simultaneously restricting the classifier to remain fair across all prospective data perturbations. Their architecture leverages an online learning algorithm (Abernethy, Hazan, and Rakhlin 2009; Shalev-Shwartz 2007) to iteratively update the model weights and data-importance profiles, guaranteeing that group constraints like DP and EO remain satisfied even when the underlying distribution is subjected to adversarial shifts. Finally, Rezaei et al. (2020) approach this challenge from the first principles of DR estimation. They formulate a zero-sum min-max game between a primary predictor and a worst-case adversarial approximator of the true label distribution, subject to feature-matching and fairness constraints. This optimization yields a robust parametric conditional distribution resembling a truncated logistic regression architecture, where the underlying truncation parameters are explicitly learned to satisfy DP and EO objectives under worst-case log-loss distribution shifts.

Complementary to these reweighting- and game-theoretic formulations, Z. Jiang et al. (2023) propose Robust Fairness Regularization (RFR), which tackles distribution shift directly through the model’s weight space rather than through sample- or label-level reweighting. RFR is grounded in a theoretical equivalence between distribution shifts, data perturbations, and model weight perturbations, and identifies sufficient conditions for fairness transferability, specifically, low prediction differences between source and target distributions for each sensitive group. Rather than solving a min-max game over reweighted samples, RFR introduces a regularization term that considers the worst-case scenario within a model weight perturbation ball, and solves the resulting inner-maximization problem with a first-order Taylor expansion, requiring only two forward and

backward propagations per update. This yields a computationally lightweight alternative to the adversarial and DR formulations above, enforcing average prediction robustness so that a fair model trained on source data remains fair as the target distribution fluctuates.

Vulnerability to Data Poisoning Fairness-aware learning algorithms are vulnerable to data poisoning attacks (*e.g.*, label flipping (Paudice, Muñoz-González, and Lupu 2018)). When training data is poisoned, standard fairness methods may misinterpret the noise as bias and over-correct, leading to severe degradation in accuracy without achieving true fairness (Roh et al. 2020). Treating poisoned data as an additional bias to be fixed rather than noise to be rejected, Roh et al. (2020) proposes an approach that extends traditional Generative Adversarial Networks (GANs) by incorporating three components: a generator, a fairness discriminator to ensure independence between predictions and sensitive attributes, and a robustness discriminator that distinguishes training predictions from a clean validation set.

Fairness Generalization Models often overfit to the fairness constraints of the training set; this is known as fairness constraint overfitting (Cotter et al. 2018). A model might perfectly satisfy DP on the training data but fail to satisfy it on the test data because it exploited specific details of the training sample rather than learning a generalizable fair pattern. Ferry et al. (2023) propose a framework to improve fairness generalization by ensuring that fairness constraints hold over various subsets of the training data. They define perturbation sets based on the Jaccard distance, requiring the model to be fair not just on the full dataset but on all subsets within a certain distance. To make this computationally feasible, they introduce a heuristic method that enforces fairness on a collection of random subsets generated via binary masks.

Data Augmentation and Manifold Smoothing Distinct from constraint-based penalties or distributionally robust optimization loops, an emerging line of research leverages virtual sample generation to smooth decision boundaries and enforce out-of-sample equity. Originally introduced as a domain-agnostic regularization technique, manifold mixup expands the support of empirical distributions by generating virtual training instances via linear convex combinations of random data pairs. Extending this mathematical paradigm to bias mitigation, Chuang and Mroueh (2021) propose *Fair Mixup* to resolve the common vulnerability where group fairness constraints satisfied during optimization fail to generalize at evaluation time. Rather than enforcing static penalties on absolute subgroup performance gaps, their framework regularizes the network along a continuous interpolation path of virtual samples transitioning gradually between distinct demographic cohorts. By explicitly minimizing the directional derivative and forcing smooth predictive transitions along these interpolation trajectories, *Fair Mixup* imposes a strong smoothness prior on the model class. This path-based smoothing ensures that key fairness metrics, such as DP and EO, generalize effectively to out-of-distribution test sets without inducing severe degradation in baseline predictive accuracy.

Domain Transfer and Generalization for Fairness A further research line addresses settings where the sensitive attribute itself is unreliable or entirely unavailable at deployment time, or where the deployment environment evolves over time. Islam, Chen, and Cai (2024) introduce Shared Latent Space-based Debiasing (SLSD), which addresses group fairness when protected demographic attributes are missing in the target domain, under the hypothesis that correlations learned in a labeled source domain can be transferred to an unlabeled target domain through a shared latent space. Its architecture proceeds through three stages: pre-training, fine-tuning, and debiasing. Pre-training uses deep canonical correlation analysis together with a specialized sampling technique to map source and target data into correlated representations; fine-tuning applies consistency training with Cauchy-distribution jittering to refine group estimates and enforce invariance to small input perturbations; and debiasing employs an adversarial technique in which an adversary penalizes the classifier whenever demographic group assignments can be reconstructed from its predictions. A relaxed variant, incorporates a minimal subset of true demographic labels when available, improving estimation accuracy over the fully label-free case.

Where SLSD transfers fairness across a fixed source-target pair, Lin et al. (2023) propose Disentanglement for Counterfactual Fairness-aware Domain Generalization (DCFDG), which targets fairness generalisation across dynamic, continually evolving environments, complementing the subset-based generalisation strategy of Ferry et al. (2023) above with a representation-learning approach. DCFDG uses variational auto-encoders to partition exogenous factors into four latent variables: semantic information caused by sensitive attributes (U_s), semantic information independent of those attributes (U_{ns}), and domain-specific information both related and unrelated to the sensitive features (U_{v1}, U_{v2}). By disentangling these variables, the model strips environment-dependent and sensitive information from its classification features, relying only on invariant semantic information for final decisions. A dedicated fairness regularisation term is integrated into the evidence lower bound objective to minimise the difference in prediction probabilities under counterfactual interventions on the sensitive attribute, extending counterfactual fairness to the domain-generalisation setting.

Local Fairness via Distributionally Robust Optimization While the methods discussed above pursue global fairness under shift, a related but distinct problem is local fairness: ensuring a predictor is unbiased within every subregion of the feature space, rather than only on average. Grari et al. (2024) propose robust optimisation for adversarial debiasing, which integrates DR optimisation into a fair adversarial objective to target this problem, employing an instance-level reweighting strategy that prioritises feature regions where the adversary faces the least difficulty reconstructing sensitive attributes. This connects the sample-reweighting logic of Du and Wu (2021) and Mandal et al. (2020) to the finer granularity of subregion-level fairness. The method is presented in two forms: a non-parametric variant with an analytical solution based on an inverted Boltzmann distribution, and a parametric version that uses a neural network to impose locality smoothness. The parametric implementation ensures that similar individuals receive similar attention from the

reweighting mechanism, preventing the model from neglecting sparsely populated or highly biased segments of the population.

Distribution-Shift Resilience in Tabular Learning Beyond methods explicitly formulated around fairness, a broader body of work addresses robustness to distribution shift in tabular learning generally and offers methodological building blocks relevant to fairness-aware tabular pipelines. Ying et al. (2025) propose shift-aware feature transformation, which addresses the deterioration of automated feature engineering under distribution shift by reframing tabular learning as a continuous representation-generation paradigm rather than a discrete search over feature sets. These models transformed features as dynamic feature-feature interaction graphs and use gradient-based optimisation to identify optimal transformations, achieving robustness through three integrated mechanisms: shift-resistant representation via embedding decorrelation and sample reweighting; flatness-aware generation that aggregates suboptimal embeddings to locate stable regions of the loss landscape, and normalisation-based alignment to bridge distribution gaps between training and test data.

Robustness Bias (RB) There is a geometric tension between adversarial robustness and recourse-based fairness, where margin maximization robustness requires pushing data points as far as possible from the decision boundary, enforcing large margins. However, recourse, the ability to change a negative outcome to positive, requires the boundary to be accessible. If a model maximizes robustness, it may make recourse disproportionately difficult for different groups, specifically for those who are situated further from the boundary. This concept is known as **RB**, Nanda et al. (2021) argue that a model is unfair if it is significantly easier to attack one demographic group than another, even if their accuracies are identical.

Sharma et al. (2021) presents a framework called *FaiR-N*, which introduces a novel loss function that approximates the distance of data points to the decision boundary using the networks' logits. It aims to equalize the average distance to the boundary for negative instances across protected groups, ensuring equal burden of recourse, and maximizes the average distance of all points to the boundary, enhancing adversarial general robustness. This mechanism allows the model to optimize for robustness equally across groups, however, it does not guarantee any fairness constraints to ensure metrics like **DP** and **EO**.

To achieve this joint optimization without incurring intractable Hessian computations from gradient-based geometric distance formulas, *FaiR-N* (Sharma et al. 2021) operates on a local linearity assumption near the decision. It approximates the distance of an instance x to the decision boundary $\mathcal{B}$ using the absolute difference between its raw network logits:

$$\hat{d}(x, \mathcal{B}) = |g_0(x) - g_1(x)|, \quad (2.9)$$

where $g_0(x)$ and $g_1(x)$ represent the pre-softmax inputs (logits) for the negative and positive classes, respectively. Using this distance proxy, the method formalizes two distinct structural

indices to regularize the decision space. The first is the fairness index (I_{fair}), which targets individuals who have been subjected to an unfavorable or negative prediction ($c(x) = 0$) and thus require recourse Sharma et al. 2021. The framework computes a demographic fairness loss as the absolute difference between the expected boundary distances of the protected groups:

$$\mathcal{L}_{\text{fairness}} = \left| \mathbb{E}_{x|s(x)=a}[\hat{d}(x, \mathcal{B})] - \mathbb{E}_{x|s(x)=b}[\hat{d}(x, \mathcal{B})] \right|, \quad (2.10)$$

where $s(x)$ partitions the data into demographic subgroups a and b . This loss is subsequently mapped to an exponential recourse-based fairness index, $I_{\text{fair}} = e^{-\mathcal{L}_{\text{fairness}}}$. Where a perfectly equitable classifier achieving an identical average burden of recourse across groups yields an index value of 1. Concurrently, the framework also tracks global model stability using a robustness index (I_{robust}). Unlike the fairness penalty, this metric averages the boundary distance across all individuals in the training cohort, regardless of their predicted outcome or demographic status. Sharma et al. (2021) defines $I_{\text{robust}} = \mathbb{E}[\hat{d}(x, \mathcal{B})]$, where maximizing I_{robust} forces the decision boundary to maintain a larger margin from all data clusters, reducing the likelihood that minor input perturbations will flip a sample's classification.

Ultimately, *FaiR-N* integrates these components with standard empirical risk minimization by formulating a combined objective function. The overall training loss $\mathcal{L}_{\text{overall}}$ is defined as:

2.4 Evolution of Continuous Generative Modeling

While the previous sections establish the necessity of jointly optimizing fairness and robustness, they reveal that standard regularization approaches either enforce rigid, context-agnostic fairness constraints that can collapse under distribution shift or deploy unconstrained manifold smoothing that floods the feature space with potential geometrically incoherent synthetic instances. This motivates the need for synthetic data generation in a principled manner that preserves geometric structure while simultaneously reinforcing decision boundaries. The answer lies in continuous generative modeling.

The evolution of generative modeling has seen a paradigm shift away from static, discrete-layer architectures toward continuous-time variants formalized by Neural Ordinary Differential Equation (ODE)s (Chen et al. 2018). Traditional Normalizing Flows define a discrete sequence of invertible transformations designed to map a simple base distribution (e.g., a standard isotropic Gaussian) to a complex, high-dimensional target data distribution. Continuous Normalizing Flow (CNF) architectures extend this continuous concept by replacing discrete layers with a continuous time variable $t \in [0, 1]$. In this continuous paradigm, the dynamic mapping of data coordinates is governed by a time-varying vector field $u_t(x)$, and the resulting temporal evolution of the time-dependent data density p_t is characterized mathematically by the continuity equation:

$$\frac{\partial p_t(x)}{\partial t} = -\nabla \cdot (p_t(x)u_t(x)). \quad (2.11)$$

Table 2.3: Paradigms at the Intersection of Fairness and Robustness.

Core Algorithmic Framework	Operational Mechanism	Challenge & Fairness/Robustness Target
Min-Max DR optimization & On-line Reweighting (Du and Wu 2021; Mandal et al. 2020)	Minimizes worst-case training loss over a structured Wasserstein ball of distribution perturbations.	Sample Selection Bias: Prevents group constraints (DP , EO) from degrading under out-of-distribution shifts.
Fair Mixup (Chuang and Mroueh 2021)	Generates virtual training paths via convex combinations transitioning between demographic cohorts.	Fairness Generalization: Minimizes directional derivatives along interpolation paths to enforce fairness metrics at test time.
FR-Train / Tripartite GAN (Roh et al. 2020)	Deploys a clean validation set monitored by a dedicated robustness discriminator alongside a fairness discriminator.	Data Poisoning: Distinguishes malicious label-flipping noise from historical demographic bias to prevent over-correction.
Jaccard Perturbation Masking (Ferry et al. 2023)	Enforces strict fairness criteria over an ensemble of stochastically masked training subsets.	Constraint Overfitting: Stops the model from exploiting high-frequency sampling artifacts unique to training.
FaiR-N Architecture	Logit proxy: $\hat{d}(x, \mathcal{B}) = g_0(x) - g_1(x) $ Loss: $\mathcal{L}_{\text{overall}} = \mathcal{L}_{\text{cross}} + \lambda_F \mathcal{L}_{\text{fairness}} + \lambda_R \frac{1}{I_{\text{robust}}}$	Robustness Bias: Harmonizes the average geometric burden of recourse across groups while maximizing classification margins.
SLSD / R-SLSD (Islam, Chen, and Cai 2024)	Deep CCA pre-training aligns domains into a shared latent space; consistency training with Cauchy jittering refines group estimates; adversary penalizes group label reconstruction.	Missing Demographics: Transfers group fairness from labeled source to unlabeled target domain.
Robust Fairness Regularization (RFR) (Z. Jiang et al. 2023)	Regularizes over worst-case model weight perturbation within a bounded ball via first-order Taylor expansion.	Distribution Shift Fairness: Preserves demographic parity when target distribution diverges from source.
DCFDG (Lin et al. 2023)	variational auto-encoder disentangles factors into sensitive, non-sensitive, and domain-specific latent variables; fairness term added to ELBO minimizes counterfactual prediction differences.	Domain Generalization: Maintains counterfactual fairness across continually evolving, unseen environments.
ROAD / BROAD (Grari et al. 2024)	Distributionally robust adversarial objective with instance-level reweighting; BROAD: analytical inverted-Boltzmann solution; ROAD: neural network with locality smoothness.	Local Fairness: Ensures unbiased predictions within every feature-space subregion.
Shift-Aware Feature Transformation (SAFT) (Ying et al. 2025)	Models transformed features as dynamic feature-feature interaction graphs; combines embedding decorrelation, flatness-aware generation, and normalization-based alignment.	Feature Drift: Preserves automated feature engineering quality under covariate shift.

$$\mathcal{L}_{\text{overall}} = \mathcal{L}_{\text{cross}} + \lambda_F \cdot \mathcal{L}_{\text{fairness}} + \lambda_R \cdot \frac{1}{I_{\text{robust}}}, \quad (2.12)$$

where $\mathcal{L}_{\text{cross}}$ represents the standard cross-entropy loss. The terms λ_F and λ_R serve as explicit, tunable hyperparameters that allow practitioners to navigate the direct architectural trade-offs between predictive accuracy, group-level recourse parity, and local adversarial robustness.

To synthesize novel data samples or evaluate exact probability densities, **CNFs** rely on numerical **ODE** solvers to integrate the learned vector field forward or backward across the temporal horizon. However, optimizing **CNF** architectures via maximum likelihood estimation requires continuously simulating these macroscopic **ODE** paths during every single gradient step of the training loop to compute exact log-likelihoods. This renders standard **CNFs** computationally intractable when processing high-dimensional manifolds or massive tabular distributions (Tong, Fatras, and Bengio 2023).

2.4.1 Flow Matching: A Simulation-Free Paradigm

To resolve the simulation constraints that paralyzed continuous normalizing flows, **FM** was introduced as a simulation-free training paradigm (Lipman et al. 2022; Albergo and Vanden-Eijnden 2022). Rather than running costly numerical **ODE** integrations to calculate likelihood gradients, **FM** reframes vector field optimization as an efficient regression task. The objective focuses on training a parameterized neural vector field, $v_\theta(t, x)$, to directly mimic the true, underlying marginal vector field, $u_t(x)$, that generates the continuous transformation from the source distribution toward the target distribution.

The standard unconstrained **FM** objective function is formalized as minimizing the expected L_2 distance between the parameterized neural network and the true target vector field:

$$\mathcal{L}_{\text{FM}}(\theta) := \mathbb{E}_{t \sim U(0,1), x \sim p_t(x)} \|v_\theta(t, x) - u_t(x)\|^2. \quad (2.13)$$

While mathematically elegant, the primary operational hurdle of the $\mathcal{L}_{\text{FM}}$ objective is its practical intractability when applied to arbitrary empirical datasets. The true, macro-level marginal vector field $u_t(x)$ and the corresponding time-dependent density $p_t(x)$ are fundamentally unknown. Consequently, they cannot be evaluated directly without solving the simulations that the framework is explicitly trying to avoid.

2.4.2 Conditional Flow Matching Formalization

CFM (Tong et al. 2023) provides the mathematical resolution to this intractability. By introducing a latent conditioning variable z , **CFM** shifts the optimization problem away from modeling the global, unconstrained flow of the entire dataset toward regressing against simple, highly localized probability paths constructed between individual data points.

As established by Tong, Fatras, and Bengio (2023), the intractable global marginal vector field $u_t(x)$ can be reconstructed as a linear mixture of simple conditional probability paths $p_t(x | z)$ and

conditional vector fields $u_t(x | z)$, integrated across the support of the conditioning distribution $q(z)$:

$$u_t(x) := \mathbb{E}_{q(z)} \left[\frac{u_t(x | z) p_t(x | z)}{p_t(x)} \right]. \quad (2.14)$$

By defining the conditioning variable z as specific pairs of source and target coordinates ($z = (x_{\text{src}}, x_{\text{tgt}})$), the optimization objective is successfully transformed into the tractable **CFM** loss formulation:

$$\mathcal{L}_{\text{CFM}}(\theta) := \mathbb{E}_{t \sim U(0,1), z \sim q(z), x \sim p_t(x|z)} \|v_\theta(t, x) - u_t(x | z)\|^2. \quad (2.15)$$

Crucially, the analytical gradients of the intractable marginal **FM** loss and the tractable **CFM** loss are mathematically identical ($\nabla_\theta \mathcal{L}_{\text{FM}} = \nabla_\theta \mathcal{L}_{\text{CFM}}$). This foundational equivalence guarantees that by optimizing a neural network to regress against simple, straight-line conditional paths between localized data pairs, the network inherently recovers the highly complex, global vector field of the dataset, bypassing the necessity for online **ODE** simulations during the training phase.

2.4.3 Optimal Transport and Path Straightening

The final theoretical challenge encountered in continuous flow modeling relates to the regression target’s variance. If the conditioning pairs $z = (x_{\text{src}}, x_{\text{tgt}})$ are sampled stochastically and independently from the marginal distributions, a configuration known as Independent Conditional Flow Matching (**I-CFM**), the resulting localized probability trajectories routinely intersect. These crossing paths force the parameterized vector field to model highly complex and curved velocity dynamics. This structural complexity significantly slows down neural network optimization convergence and degrades out-of-sample generation quality.

OT-CFM directly eliminates this trajectory variance by enforcing a joint pairing distribution $q(z)$ that is explicitly aligned with the 2-Wasserstein **OT** plan. By solving a discrete linear assignment problem over the data cohorts, **OT-CFM** couples each source coordinate uniquely with its geometrically closest counterpart in the target distribution, minimizing the global quadratic displacement cost:

$$\min_{\pi} \int \|x_{\text{src}} - x_{\text{tgt}}\|^2 d\pi(x_{\text{src}}, x_{\text{tgt}}). \quad (2.16)$$

This cost-minimizing alignment guarantees mathematically straight, monotonic, and non-crossing displacement trajectories across the continuous time horizon. Not only does this path-straightening effect accelerate the training of neural vector fields, but it also establishes a stable, deterministic geometric mapping between disparate domains.

2.5 Final Remarks

The convergence of the foundational pillars detailed in this literature review, statistical group equity, margin-based boundary resilience, intersectional optimization pathologies, and optimal transport flow dynamics, reveals under-addressed gaps in contemporary **ML** research.

While in-processing fairness constraints have proliferated, their empirical behavior and failure modes under severe class imbalance remain incompletely understood. Existing frameworks are not systematically evaluated across imbalanced domains or under performative distribution shifts, leaving a critical gap in our understanding of how these methods degrade in realistic, skewed data environments. Current literature lacks comprehensive empirical auditing frameworks that track fairness and robustness trajectories over extended deployment horizons on imbalanced datasets.

Manifold smoothing techniques like *Fair Mixup* improve fairness generalization by generating virtual training instances, but they operate without explicit geometric constraints. These methods flood the feature space with unconstrained convex combinations that may violate the structure of the underlying data manifold, leading to spatially inconsistent decision boundaries under performative pressure. The literature lacks principled frameworks that integrate synthetic data generation with geometric coherence guarantees or boundary-aware sampling strategies.

Although optimal transport theory has revolutionized generative modeling through **CFM** and path-straightening mechanisms, its theoretical and practical integration with fairness optimization remains largely unexplored. Existing fairness-robustness frameworks do not leverage the deterministic, non-crossing trajectory properties that **OT-CFM** provides, missing an opportunity to enforce geometric structure during augmentation.

The intersection of fairness and robustness has been studied primarily in static, adversarial, or naturally drifting settings. However, performative feedback loops, where algorithmic predictions actively reshape the underlying relations between features and output, introduce a distinct class of challenges that remain largely absent from fairness-robustness literature. The behavior and stability of fairness-aware classifiers under performative distribution shifts has received limited empirical attention, leaving open questions about how and why these methods may degrade when predictions influence the data itself.

These gaps collectively suggest that a more integrated approach is needed, one that combines principled synthetic generation with explicit boundary reinforcement, leverages optimal transport geometry for structural coherence, and evaluates fairness-robustness trade-offs under realistic, dynamic deployment scenarios rather than static benchmarks.

Chapter 3

FairMorph

This chapter introduces our primary methodological contribution, detailing the development and execution of a morphing-based data augmentation framework designed to systematically mitigate demographic robustness bias in tabular classifiers.

To map out this architecture, the chapter is organized as follows:

- **Section 3.1: Problem Scope** establishes the foundational boundaries of our environment, formalizing the operational challenges, data-dependent constraints, and real-world implications of algorithmic unfairness in high-stakes domains.
- **Section 3.2: Dataset Morphing via Optimal Transport** provides the underlying geometric foundation of our approach, outlining how cost-minimizing matching bijections are solved to map cross-cohort neighborhoods without relying on heavy generative models.
- **Section 3.3: Dynamic Augmented Training for Fair Robustness** introduces our initial exploratory paradigm, detailing how the framework constructs fairer training batches on-the-fly and analyzing the structural trade-offs of distinct convex-hull sampling topologies.
- **Section 3.4: FairMorph** details our definitive proposal, formalizing the targeted, conditional adversarial pipeline that leverages the continuously evolving intermediate classifier to reinforce fragile decision boundaries while completely eliminating data redundancy.

3.1 Problem Scope

ML has been increasingly used in sensitive domains, including criminal justice (Stevenson 2018), healthcare (Rajkomar, Dean, and Kohane 2019), and financial services (Kaur, Kumar, and Kaur 2023). These systems are asked to perform high-impact decisions, often without human supervision. At the same time, ML models are usually optimized for a single global metric (e.g., empirical risk minimization via binary cross-entropy loss), without taking into account localized sub-group dynamics and their real-world implications.

A critical side-effect of this global optimization paradigm is the emergence of robustness bias, which manifests as highly disparate classification error rates and boundary fragility across

sub-groups of individuals based on sensitive attributes such as ethnicity, skin color, age, or gender (Barocas and Selbst 2016; Bolukbasi et al. 2016; Gonen and Goldberg 2019). Geometrically, because the majority population dominates the loss function, the classifier prioritizes maximizing the decision space around majority data clusters. Furthermore, the resulting decision boundary is pushed disproportionately close to minority instances. This creates disproportionate margins across groups, leaving the minority demographic with a narrow, highly volatile classification margin. Under these conditions, minor input perturbations or temporal data shifts easily flip minority predictions, creating localized fragility that traditional global metrics completely fail to detect. While historical data biases, under-represented populations, or label noise compound this issue (Suresh and Guttag 2019), the fundamental geometric nature of learning under class-imbalance inherently trades minority robustness for majority accuracy.

Algorithmic fairness is an emerging sub-field in ML aiming at studying and mitigating discrimination in the algorithmic decision-making process. However, current bias mitigation approaches suffer from structural limitations, routinely acting on sub-optimal spaces or failing to address the underlying margin disparities. More specifically, we argue that pre-processing methods, by altering the data distribution before the learning phase, cannot explicitly guarantee or control how the final, complex classifier will position its boundaries in production. While post-processing methods attempt to shift classification thresholds after a model has been learned. This can be viewed as knowingly training a geometrically biased model and then applying a superficial correction, which adds systemic complexity and leaves the underlying fragile margins unaddressed instead of learning an equitable boundary from the outset. On the other hand, traditional in-processing methods modify the objective function directly but are distinctly model-specific and metric-specific, possibly hindering architectural flexibility and real-world usefulness. Crucially, none of these paradigms dynamically manage the structural distribution of margins across demographic manifolds during the active training process.

The current landscape of algorithmic fairness lacks practical, domain-agnostic methodologies for seamless real-world deployment. While regulatory frameworks such as the GDPR increasingly mandate strict algorithmic accountability and fairness, translating these legal requirements into production-ready pipelines introduces significant engineering friction. ML practitioners frequently face the complex task of balancing mandatory legal compliance with the pragmatic risks of model performance degradation and development overhead. To achieve a meaningful impact in our communities, bias mitigation must be computationally efficient, work out-of-the-box, and seamlessly integrate into the standard pipelines that enterprises already employ. By introducing fairness-awareness directly into the active training phase via geometric regularization, we can enable organizations to accommodate equity with minimal friction and little to no extra operational cost.

Concretely, these high-level objectives translate into the following methodological goals, which guide the design of *FairMorph*:

- 1. Empirically study the fairness-robustness relation, explicitly analyzing how dynamic data augmentation can be leveraged to navigate and optimize the trade-off between global

accuracy and sub-group boundary stability.

- 2. Develop a dynamic, geometry-aware augmentation method that identifies low-margin, vulnerable, or low-density sub-groups within the input space on-the-fly, utilizing **OT** to systematically expand minority decision margins.
- 3. Evaluate performance within the robustness-bias landscape, comparing our framework against traditional margin optimization techniques to demonstrate its capacity to mitigate overconfidence and protect minority demographics from performative fairness degradation.

3.2 Dataset Morphing via Optimal Transport

The concept of dataset morphing, initially proposed by Correia, Soares, and Jorge (2019), involves generating intermediate datasets that gradually and continuously transition from a source to a target distribution. In this work, we adapt dataset morphing as a foundational geometric data augmentation strategy. By leveraging deterministic mathematical interpolation, the framework samples the underlying manifold between disparate domains, generating synthetic tabular instances that exhibit a controlled, parameterized mixture of properties from both distributions.

Rather than relying on computationally intensive generative neural networks (e.g., GANs, VAEs) or complex neural **ODE** solvers, our framework employs an efficient, geometric approach: discrete **OT** coupled with linear displacement interpolation.

Operationally, this foundational framework implements morphing via a two-stage pipeline. First, to ensure structurally meaningful morphological transitions, the algorithm must establish optimal pairings between individual samples from the source distribution $\mathcal{D}_{\text{src}}$ and the target reference distribution $\mathcal{D}_{\text{tgt}}$. To avoid intersecting trajectories and preserve the underlying data geometry, we solve for the matching configuration that minimizes the global displacement cost. This is formulated as a discrete **OT** problem, where we compute a pairwise cost matrix C based on the squared Euclidean distance between individual feature vectors:

$$C_{i,j} = \|x_i - x_j\|^2 \quad \text{where} \quad x_i \in \mathcal{D}_{\text{src}}, x_j \in \mathcal{D}_{\text{tgt}}. \quad (3.1)$$

The optimal one-to-one assignment is then solved using the Hungarian algorithm (Kuhn 1955), which determines a cost-minimizing bijection π such that each vulnerable source sample x_s is uniquely mapped to its structurally closest target anchor $x_t = x_{\pi(s)}$.

Second, to construct the augmented feature spaces, we define a continuous interpolation function τ that governs the displacement trajectory across a continuous coefficient $\alpha \in [0, 1]$. Because the **OT** assignment guarantees non-crossing pairs, we can bypass the velocity vector fields and compute intermediate states using direct linear interpolation:

$$\tau(x_s, x_t, \alpha) = (1 - \alpha)x_s + \alpha x_t. \quad (3.2)$$

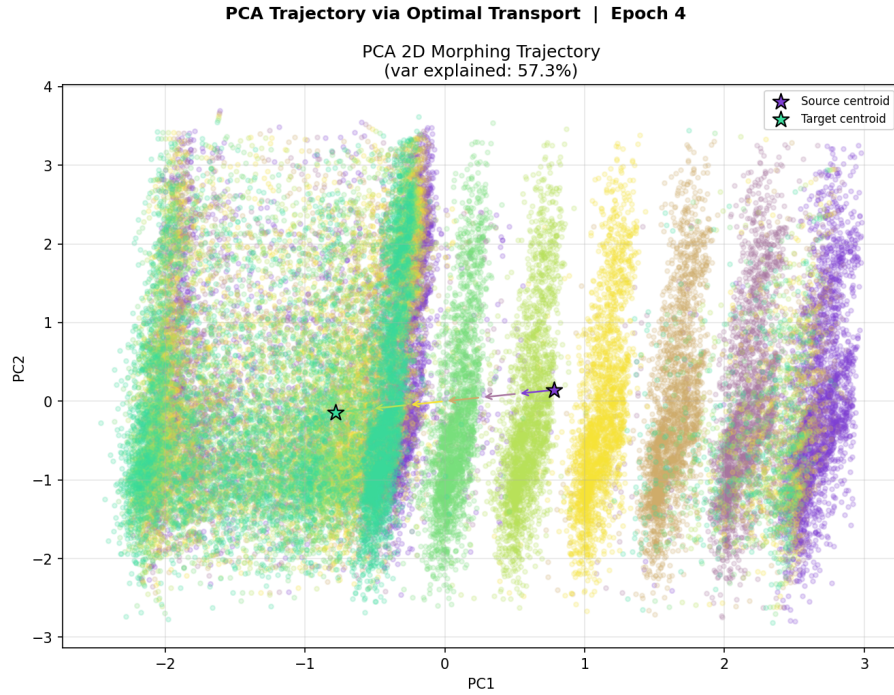


Figure 3.1: Dataset generation process via linear optimal transport mapping on the Folktable Dataset. The transformation shifts the source dataset (purple) toward the target dataset (cyan), with stars marking respective centroids and arrows indicating the direction of optimal transport mapping.

By evaluating this foundational mapping function τ at discrete intervals along the path (e.g., $\alpha \in \{0.0, \dots, 1.0\}$), the pipeline extracts intermediate snapshots that seamlessly bridge the source and target domains. This displacement function serves as the core generative engine for both our exploratory batch sampling topologies and our targeted adversarial threshold optimization loops. Figure 3.1 showcases the trajectory of the generative process, from the source samples in purple to the target data in cyan, for the Folktable Dataset (Ding et al. 2021). The calculated transport paths are represented by directional vectors tracking from the source centroid to the target centroid.

3.3 Dynamic Augmented Training for Fair Robustness

As discussed in Section 2.3, the intersection of fairness and robustness manifests in various distinct structural configurations. A critical area of research focuses on how a model’s decision boundary behaves when evaluated across disparate demographic groups. When the sensitivity or geometric stability of a model differs significantly between these groups, the system exhibits **RB**. In high-stakes enterprise domains such as financial fraud detection or credit scoring, this vulnerability translates directly into severe real-world harms, resulting in elevated rates of unjustly declined transactions or discriminatory loan defaults for the disadvantaged group.

The main objective of this first exploratory methodology is to construct an inline data augmentation mechanism that spatially interpolates the sparse regions, separating distinct group manifolds

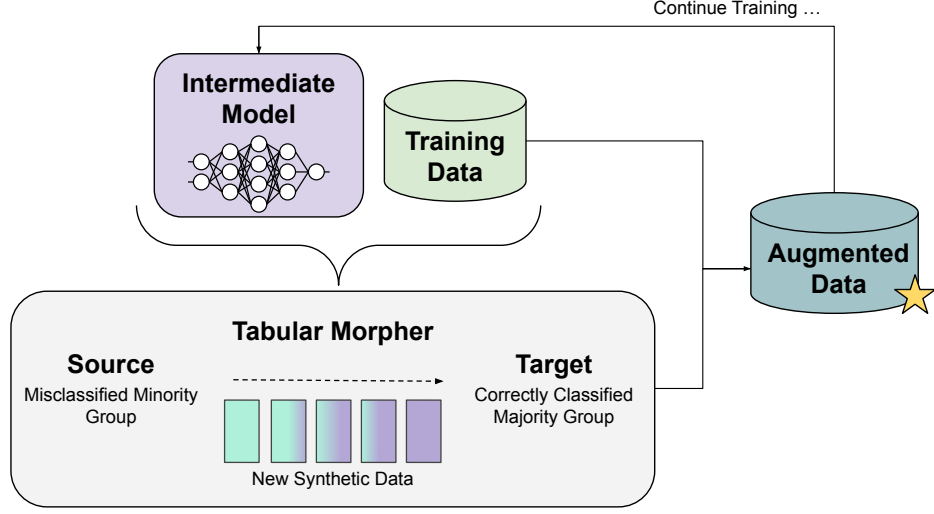


Figure 3.2: Dynamic data augmentation pipeline, showcasing the generation of new synthetic data derived from Tabular Morpher interpolation between the vulnerable source and robust target datasets.

in the feature space. By systematically populating these low-density boundary regions, we aim to mitigate the geometric separability of demographic cohorts based on sensitive attributes, thereby minimizing the model’s reliance on group-specific features when establishing its decision boundary. In contrast to traditional static data augmentation approaches that generate synthetic instances *a priori*, this framework introduces a dynamic augmentation paradigm. As illustrated in the initial design pipeline shown in Figure 3.2, the framework actively constructs fairer training batches on-the-fly, continuously adapting to the model’s evolving structural weaknesses throughout the training phase.

Formally, let each instance in the empirical dataset $\mathcal{D}$ be parameterized by the tuple (x, a, y) , where $x \in \mathcal{X}$ is the feature vector, $y \in \mathcal{Y}$ is the target label, and $a \in \{0, 1\}$ is the binary sensitive attribute ($a = 1$ denoting the disfavored group and $a = 0$ representing the majority baseline). Let f_{θ_i} denote the intermediate classifier parameterized by weights θ at training epoch i , producing a predicted label $\hat{y} = f_{\theta_i}(x)$. At each epoch, we dynamically evaluate these predictions to construct two specific subsets that form the operational basis of our augmentation:

- **Source Set** ($\mathcal{D}_{\text{src}}$): The subset of misclassified instances belonging exclusively to the disfavored minority cohort, defined as:

$$\mathcal{D}_{\text{src}} = \{(x_i, y_i) \mid (x_i, a_i, y_i) \in \mathcal{D} \wedge a_i = 1 \wedge \hat{y}_i \neq y_i\}. \quad (3.3)$$

- **Target Set** ($\mathcal{D}_{\text{tgt}}$): The subset of correctly classified instances belonging to the majority baseline cohort, defined as:

$$\mathcal{D}_{\text{tgt}} = \{(x_j, y_j) \mid (x_j, a_j, y_j) \in \mathcal{D} \wedge a_j = 0 \wedge \hat{y}_j = y_j\}. \quad (3.4)$$

By leveraging the **OT** mapping defined in the previous section, we establish a continuous structural flow from the vulnerable samples in $\mathcal{D}_{\text{src}}$ directly to the robust anchor samples in $\mathcal{D}_{\text{tgt}}$. By sampling across this interpolation continuum, the pipeline generates K intermediate morphed datasets. Injecting these intermediate snapshots back into the optimization loop forces the next state of the classifier, $f_{\theta_{i+1}}$, to learn a more equitable and robust decision boundary, effectively mapping out and correcting the spatial blind spots where the minority group previously failed.

Synthetic Data Sampling Topologies

Because generating intermediate snapshots across matched pairs can yield an expansion of $K|\mathcal{D}_{\text{src}}|$ synthetic instances, feeding this unconstrained data volume directly into the training loop can drastically destabilize classifier behavior and degrade predictive performance. To control this expansion, we enforce a data budget by undersampling the synthetic pool to a maximum ceiling of $|\mathcal{D}_{\text{src}}|$. Consequently, when augmenting empirical tabular distributions with morphed snapshot sets, the specific downsampling strategy applied across these paths directly governs the structural composition and regularizing impact of the active training batches.

We explore this generative dynamic across three distinct distribution topologies, exposing an operational trade-off between semantic data fidelity and geometric regularization strength. Because our framework constructs synthetic samples via linear interpolation between paired coordinates, the generated instances are mathematically guaranteed to reside within the continuous convex hull of the empirical data. However, real-world tabular data manifolds, particularly in highly sensitive financial or census environments, are inherently non-convex, characterized by non-linear feature dependencies and low-density areas. Consequently, drawing samples across a continuous convex set introduces a trade-off regarding data fidelity and behavioral realism:

- **Uniform Random Sampling:** This topology serves as a baseline control by drawing synthetic samples across all discrete interpolation steps with equal probability. While statistically unbiased, it spreads its generative weight evenly across the entire convex path, potentially diluting its localized impact on both group fairness metrics and boundary margin optimization.
- **Hill-Shaped (Centric) Sampling:** This strategy concentrates generative weight heavily around the midpoint snapshots ($t \approx 0.5$). Geometrically, this approach populates the lowest-density, most ambiguous intersection zones between demographic groups. While highly effective at optimizing fairness metrics by dismantling rigid, biased decision surfaces, it exhibits lower data fidelity; generating hybrid feature combinations far from empirical data clusters forces the model to learn from less realistic instances, which can slightly degrade global predictive accuracy.
- **Valley-Shaped (Extremity) Sampling:** This approach skews sample selection toward the boundaries of the interpolation path (nearest to $t = 0.0$ and $t = 1.0$). By staying close to the original source and target manifolds, this topology prioritizes structurally realistic samples

that maximize data fidelity. While this maintains baseline model utility and accuracy, it fails to adequately reshape the underlying boundary, leaving demographic disparities undressed or slightly exacerbated.

3.4 FairMorph: Adversarial Augmented Training

While broad batch-level sampling over the convex hull successfully regularizes the network, it incurs significant computational overhead and risks injecting redundant or low-utility synthetic data. To optimize this process, we propose a specialized paradigm that shifts focus away from continuous manifold population toward targeted adversarial boundary reinforcement.

The intuition driving this new architecture is rooted directly in the principle of demographic invariance. In an equitable algorithmic system, a fair model’s decision boundary must remain invariant to changes along a protected or sensitive attribute axis. Consequently, if two samples belonging to entirely different demographic cohorts sit close enough to one another within the underlying feature space, they should inherently share the same classification label.

By tracking these close, cross-demographic neighbor pairings, we can identify exactly where the current model is failing this invariance test. Rather than filling the global feature space blindly, our framework isolates these specific points of friction, generating targeted synthetic edge cases to explicitly force the model to harmonize its predictions across demographic lines.

To ground this intuition mathematically, let each instance in the empirical dataset $\mathcal{D}$ be parameterized by the tuple (x, a, y) , where $x \in \mathcal{X}$ is the feature vector, $y \in \mathcal{Y}$ is the target label, and $a \in \{0, 1\}$ is the binary sensitive attribute. We isolate the demographic cohorts by partitioning the global distribution into a source set $\mathcal{D}_{\text{src}}$ representing the disfavored subgroup ($a = 1$), and a target reference set $\mathcal{D}_{\text{tgt}}$ capturing the favored majority baseline population ($a = 0$):

- **Source Set** ($\mathcal{D}_{\text{src}}$): The subset of instances belonging exclusively to the disfavored minority cohort, defined as:

$$\mathcal{D}_{\text{src}} = \{(x_i, y_i) \mid (x_i, a_i, y_i) \in \mathcal{D} \wedge a_i = 1\}. \quad (3.5)$$

- **Target Set** ($\mathcal{D}_{\text{tgt}}$): The subset of instances belonging to the majority baseline cohort, defined as:

$$\mathcal{D}_{\text{tgt}} = \{(x_j, y_j) \mid (x_j, a_j, y_j) \in \mathcal{D} \wedge a_j = 0\}. \quad (3.6)$$

To map the cross-cohort neighborhoods, pairs are computed via **OT**. We define a minimum-cost coupling matrix that matches each disfavored instance in $\mathcal{D}_{\text{src}}$ with its geometrically closest counterparts in the favored majority distribution $\mathcal{D}_{\text{tgt}}$.

Crucially, this entire augmentation pipeline occurs dynamically during network training; at each given training epoch, the framework evaluates the data manifold using the intermediate state of the classifier to determine exactly which synthetic instances should be introduced into the optimization loop. This paradigm processes the **OT**-paired instances conditionally, injecting at most one strategic sample per source instance into the active training batch for that specific epoch. By

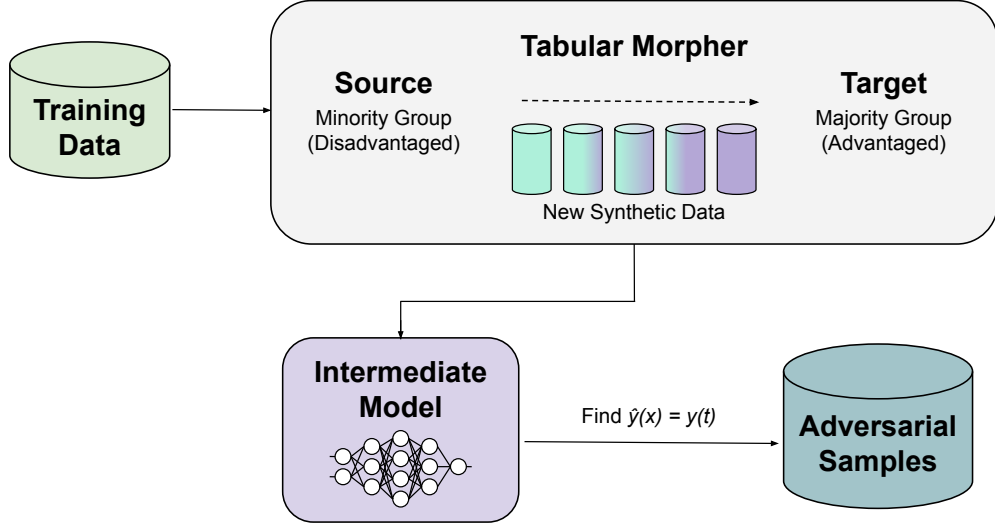


Figure 3.3: *FairMorph* data augmentation pipeline. At each epoch, the *Tabular Morpher* pairs source and target instances via OT and generates intermediate synthetic snapshots along their linear interpolation trajectory. These snapshots are passed through the intermediate classifier to identify the adversarial samples, where the prediction first flips to the target label.

design, this constraints the generation phase to a worst-case upper bound of exactly $|\mathcal{D}_{\text{src}}|$ newly synthesized samples, preventing unconstrained data expansion from destabilizing the model.

For each individual sample $(x_s, y_s) \in \mathcal{D}_{\text{src}}$ coupled with its target instance $(x_t, y_t) \in \mathcal{D}_{\text{tgt}}$, the pipeline audits their respective ground-truth labels. If the source sample already shares the same class outcome as the target destination ($y_s = y_t$), the demographic invariance condition is satisfied; no cross-class boundary mapping is required, and the pair is skipped entirely. Conversely, when the labels differ ($y_s \neq y_t$), the framework triggers the morphing engine to generate a discrete sequence of K intermediate snapshots along the interpolation trajectory:

$$\mathcal{S} = \{x_k \mid x_k = \tau(x_s, x_t), k \in \{1, \dots, K\}\}. \quad (3.7)$$

During an active training epoch, all generated snapshots in the sequence $\mathcal{S}$ are forwarded through the current intermediate classifier to compute their predicted labels $\hat{y}(x_k)$. The pipeline then filters this sequence to extract exclusively the single adversarial edge case x_{adv} where the intermediate model’s prediction first switches to match the target label:

$$x_{\text{adv}} = x_{k^*} \quad \text{where} \quad k^* = \min \{k \in \{1, \dots, K\} \mid \hat{y}(x_k) = y_t\}. \quad (3.8)$$

By injecting only this precise boundary-crossing snapshot into the active epoch’s training batch, the framework eliminates data redundancy. It uses the intermediate model’s own feedback to isolate the point of transition where the classifier’s group distinctions blur, forcing the network to narrow its demographic gap and learn a more robust, equitable decision surface without altering the overall integrity of the original data distribution.

Table 3.1: Comparison between sampling approaches and *FairMorph* augmentation averaged across multiple training phases of the Folktables dataset with the original ground truth. Values in bold highlight the best-performing method regarding accuracy, F1-score, and the ratio of fairness metrics across groups.

Method	Group	Acc $\uparrow$	F1 $\uparrow$	DP $\uparrow$	TPR $\uparrow$	FPR $\downarrow$
FairMorph	0	0.781	0.783	0.560	0.814	0.268
	1	0.781	0.783	0.381	0.724	0.157
	Ratio	-	-	0.677	0.886	0.578
Hill-shaped Sampling	0	0.779	0.779	0.559	0.810	0.270
	1	0.779	0.779	0.357	0.694	0.137
	Ratio	-	-	0.638	0.855	0.503
Random Sampling	0	0.782	0.785	0.565	0.820	0.273
	1	0.782	0.785	0.364	0.707	0.140
	Ratio	-	-	0.643	0.861	0.507
Valley-shaped Sampling	0	0.781	0.782	0.560	0.813	0.269
	1	0.781	0.782	0.358	0.699	0.135
	Ratio	-	-	0.638	0.858	0.495
Vanilla	0	0.777	0.773	0.536	0.787	0.247
	1	0.777	0.773	0.341	0.674	0.124
	Ratio	-	-	0.634	0.853	0.496

Augmentation Approaches Comparison

To evaluate the efficacy of the targeted adversarial boundary reinforcement mechanism in *FairMorph*, we benchmark it against our three exploratory convex-hull sampling topologies (Uniform Random, Hill-Shaped, and Valley-Shaped) as well as an unregularized Vanilla baseline. Each framework is evaluated across standard predictive utility metrics (Accuracy and F_1 -score) and demographic group equity metrics (DP, TP, and FP parity ratios). Experiments are conducted on the Folktables benchmark suite Ding et al. (2021) configured for the ACSIncome task, where the protective vector isolates racial demographic groups.

The empirical outcomes summarized in Table 3.1 demonstrate that the refined *FairMorph* paradigm yields superior optimization across all group fairness dimensions. Specifically, *FairMorph* achieves the highest group parity ratios across the board. Crucially, this leap in demographic fairness incurs a marginal utility trade-off; *FairMorph* maintains a high predictive accuracy of 0.781, trailing the top-performing accuracy model by a minute 0.1 percentage points.

Interestingly, the Uniform Random Sampling topology delivered the highest overall predictive performance, yielding an Accuracy of 0.782 and an F_1 -score of 0.785. This finding counters our initial exploratory hypothesis, which predicted that a completely unconstrained, uniform sampling scheme would dilute boundary regularization and degrade utility. Instead, the data suggests that distributing synthetic weight uniformly across the convex hull offers an effective mid-level smoothing effect, tracking slightly ahead of the more localized sampling configurations.

Conversely, the localized sampling topologies behaved strictly in accordance with our structural hypotheses. The Hill-Shaped topology, which concentrates synthetic injection entirely within

Table 3.2: Hybrid Metric Rankings. Ranked over the average of the fairness ratios and performance metrics for the different methods.

Method	Acc	F1	DP	TPR	FPR	Overall Rank
<i>FairMorph</i>	2.0	2.0	1.0	1.0	1.0	1.4
Hill-shaped Sampling	4.0	4.0	3.0	4.0	3.0	3.6
Random Sampling	1.0	1.0	2.0	2.0	2.0	1.6
Valley-shaped Sampling	3.0	3.0	4.0	3.0	5.0	3.6
Vanilla	5.0	5.0	5.0	5.0	4.0	4.8

the lowest-density intersection zones ($t \approx 0.5$), proved to be the most disruptive sampling configuration, degrading baseline accuracy to 0.779 and F_1 -score to 0.779. This drop confirms that aggressively populating non-convex tabular voids with continuous linear midpoint configurations introduces out-of-distribution semantic noise that can destabilize the global **NN** representation. Similarly, the Valley-Shaped topology, which limits its generation to the outer extremities near the empirical clusters, prioritized data fidelity at the expense of fairness optimization, failing to adequately shift the decision surface and even marginally lagging behind the Vanilla baseline’s group **FP** ratio by an absolute value of 0.001.

To synthesize these multidimensional trade-offs into a single diagnostic assessment, Table 3.2 details the individual metric rankings across all configurations, as well as the average global ranking across all metrics. Sweeping the top rank (Rank = 1.0) across every fairness category, *FairMorph* secures the best multi-objective position with an Overall Rank of 1.4. The Uniform Random approach follows as second-place alternative (Overall Rank = 1.6) due to its raw accuracy advantages. In contrast, the restricted Hill-Shaped and Valley-Shaped sampling configurations tie as the worst-performing active augmentation strategies (Overall Rank = 3.6), highlighting the pitfalls of unconstrained dataset inflation. Ultimately, all proposed augmentation variants maintain competitive predictive performance and provide an equity advantage over the standard Vanilla baseline, which sinks to the bottom of the evaluation suite with an Overall Rank of 4.8.

Chapter 4

TabMIST

This chapter introduces our main diagnostic contribution, *Tabular MIST*, and details its architectural implementation.

- **Section 4.1: Methodological Foundations** outlines the methodological foundations and motivates the transition of the sensitivity framework to the tabular domain.
- **Section 4.2: TabMIST** formalizes the mathematical formulation of the *TabMIST* metric and details its underlying algorithmic calculation.
- **Section 4.3: Evaluation of Demographic Robustness Bias** provides an illustrative example demonstrating how this framework is deployed as an auditing mechanism to evaluate demographic robustness bias.

4.1 Methodological Foundations

Ensuring model robustness in classification tasks remains a critical challenge, particularly when deploying machine learning systems in safety-sensitive or high-stakes domains (Balendran et al.; Perdomo et al. 2020; Beede et al. 2020). Traditional evaluation practices tend to focus on aggregate accuracy measures; however, these metrics fail to address or expose model instability under semantically meaningful input deformations (Sehwag et al. 2019). To mitigate this evaluation gap, the *tsMIST* framework was previously developed to systematically quantify classifier resilience via controlled interpolation trajectories between challenging inter-class samples.

As a natural architectural extension, we introduce *TabMIST*, a tabular model sensitivity metric that transitions this diagnostic framework from time series data to the tabular domain. While the original time series paradigm relies on elastic distance measures such as Dynamic Time Warping (DTW) to isolate candidate cross-class pairs with minimal distance, *TabMIST* partitions tabular datasets using boolean masking to separate the ground-truth classes, and the squared Euclidean distance to identify the closest candidates.

To formalize this environment, let each instance in the empirical dataset $\mathcal{D}$ be parameterized by the tuple (x, y) , where $x \in \mathcal{X}$ is the feature vector. Given a binary classification task, we

isolate the classification cohorts by partitioning the global distribution into a baseline source class distribution $\mathcal{D}_{\text{src}}$ and an opposing target class distribution $\mathcal{D}_{\text{tgt}}$:

- **Source Set** ($\mathcal{D}_{\text{src}}$): The subset of instances belonging to the label class 0:

$$\mathcal{D}_{\text{src}} = \{(x_j, y_j) \mid (x_j, y_j) \in \mathcal{D} \wedge y_i = 0\}. \quad (4.1)$$

- **Target Set** ($\mathcal{D}_{\text{tgt}}$): The subset of instances belonging to the label class 1:

$$\mathcal{D}_{\text{tgt}} = \{(x_j, y_j) \mid (x_j, y_j) \in \mathcal{D} \wedge y_i = 1\}. \quad (4.2)$$

Pairs are subsequently mapped out geometrically between these two subsets, and an explicit linear dataset morpher is executed to yield a smooth, sequential space of n intermediate tabular datasets, denoted as $X^{(\text{morph})}$.

By transitioning to this tabular paradigm, the framework shifts from tracking temporal signal deformations to mapping the micro-level topography of the tabular decision surface. In the following subsections, we formalize the exact algorithmic computation of the *Switch Threshold*, the precise coordinate along the interpolation trajectory where a sample’s predicted classification flips. We then detail how these localized transitions are aggregated into two macro-level structural descriptors: $\text{TabMIST}_{\text{Avg}}$, which diagnoses global boundary rigidity or fragility, and $\text{TabMIST}_{\text{Std}}$, which maps spatial consistency across the data manifold. Finally, we provide a geometric interpretation of these metrics by projecting the data manifold using Principal Component Analysis (PCA) to isolate and expose the spatial characteristics of weak and stubborn data clusters.

4.2 TabMIST: Morphing-based Sensitivity Metric

Building on the sensitivity principle introduced by Brito et al. (2025) for time series, we measure the precise structural deformation required to flip the model’s prediction from its native source class to the opposing target class. This boundary transition is mathematically captured by the *Switch Threshold*:

$$\text{Switch Threshold} = \min_{1 \leq i \leq n} i \quad \text{s.t.} \quad \hat{Y}_i^{(\text{morph})} \neq Y^{(\text{source})}, \quad (4.3)$$

where i represents the specific step index along the $X^{(\text{morph})}$ trajectory that triggers a modification in the classifier’s binary prediction. Setting $n = 100$ steps allows the threshold to be intuitively interpreted as a percentage of transformation. Across δ distinct evaluation pairs, the global robustness distribution is aggregated into two core metrics, $\text{TabMIST}_{\text{Avg}}$ and $\text{TabMIST}_{\text{Std}}$:

$$\text{TabMIST}_{\text{Avg}} = \frac{1}{\delta} \sum_{k=1}^{\delta} \text{Switch Threshold}_k, \quad (4.4)$$

Algorithm 1 TabMIST Sensitivity Metric Calculation

Require: Source dataframe $\mathcal{D}_{src}$, Target dataframe $\mathcal{D}_{tgt}$
Require: Trained classifier model f_θ
Require: Total morphing steps n
Ensure: TabMIST evaluation summary: $TabMIST_{Avg}$, $TabMIST_{Std}$, and raw scores $\mathcal{S}$

- 1: $\mathcal{X}_0, \mathcal{X}_1, \dots, \mathcal{X}_n \leftarrow \text{Fit_Morph_Linear}(\mathcal{D}_{src}, \mathcal{D}_{tgt}, n)$ ▷ Generate trajectory snapshots
- 2: $\mathcal{S} \leftarrow \emptyset$
- 3: **for** each sample row i from 1 to $N_{samples}$ **do**
- 4: Let $P_i = [f_\theta(\mathcal{X}_0[i]), f_\theta(\mathcal{X}_1[i]), \dots, f_\theta(\mathcal{X}_n[i])]$ ▷ Get prediction trajectory for sample i
- 5: **if** $P_{i,0} == 0$ **and** $P_{i,n} == 1$ **then** ▷ Filter valid transitions
- 6: $s_{flip} \leftarrow \min\{j \mid P_{i,j} == 1\}$ ▷ Find index of first prediction flip
- 7: Score $S_i \leftarrow \left(\frac{s_{flip}}{n}\right) \times 100$ ▷ Scale score to a percentage
- 8: Append S_i to $\mathcal{S}$
- 9: **end if**
- 10: **end for**
- 11: $TabMIST_{Avg} \leftarrow \text{Mean}(\mathcal{S})$
- 12: $TabMIST_{Std} \leftarrow \text{Standard_Deviation}(\mathcal{S})$
- 13: **return** $\{TabMIST_{Avg}, TabMIST_{Std}, \mathcal{S}\}$

$$TabMIST_{Std} = \sqrt{\frac{1}{\delta} \sum_{k=1}^{\delta} (Switch\ Threshold_k - TabMIST_{Avg})^2}. \quad (4.5)$$

The resulting statistical descriptors allow for a comprehensive auditing of the underlying tabular decision surface. A $TabMIST_{Avg}$ value hovering near 50% indicates optimal boundary stability, meaning that the model maintains highly stable predictions and changes its classification decision only when the underlying features have undergone substantial structural transformation toward the opposing class.

Conversely, extreme variations in the average signal reveal critical structural limitations: a significantly low $TabMIST_{Avg}$ metric implies boundary fragility where nominal feature adjustments trigger premature prediction switches, whereas an excessively high $TabMIST_{Avg}$ value points to model rigidity, where the system fails to adapt its predictions even when major target-class attributes are dominant. Furthermore, the variance of these transitions yields insight into spatial consistency across the data manifold; a low $TabMIST_{Std}$ indicates a uniform response to perturbations across the entire feature space, while a high $TabMIST_{Std}$ uncovers region-dependent vulnerabilities, where specific localized areas of the decision surface that are uniquely fragile to distribution shifts.

To easily interpret and visualize where these edge cases lie across the underlying data manifold, the feature space can be projected using **PCA**, as illustrated in Figure 4.1. This layout maps the source class (purple clusters) against the target class (cyan clusters), offering a more intuitive visualization of problematic samples where the model may struggle:

- **Stubborn Samples (Red Circles):** These instances display exceptionally high *switch thresholds*, indicating a severe resistance to reclassification. Spatially, they lie deep within the

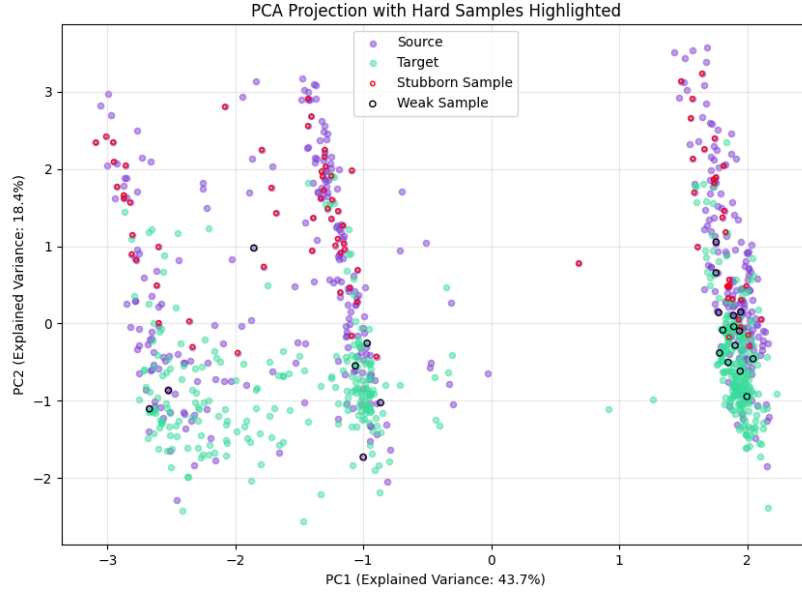


Figure 4.1: PCA projection of the data manifold highlighting the geometric positioning of stubborn (red circles) and weak (black circles) samples relative to the source and target clusters.

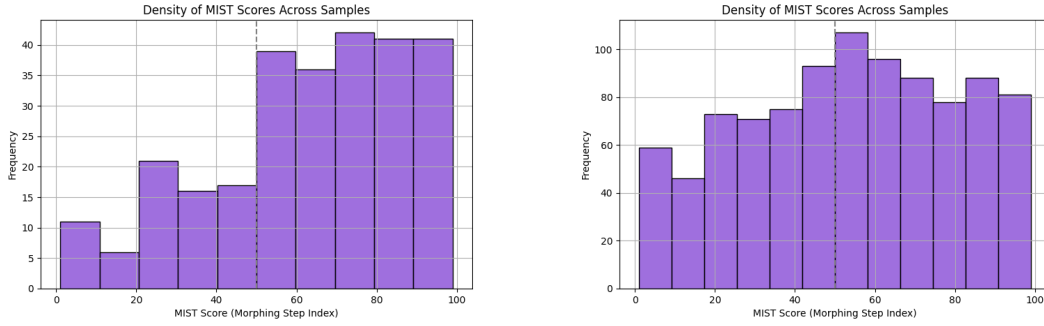
core source clusters, far from the target manifold. Because they are securely anchored inside dense source regions, the model requires more interpolation changes before shifting its prediction.

- **Weak Samples (Black Circles):** These instances represent boundary fragility with very low *switch thresholds*. Although they technically belong to the source class, they are spatially misplaced and sit directly within clusters dominated by the target class. This high exposure to the opposing distribution explains why even minor feature updates immediately trigger a prediction switch.

To provide an explicit blueprint of the robustness auditing phase, Algorithm 1 details the exact procedure for computing the *TabMIST* metric scores across a trained model space. This algorithm processes the compiled, matched evaluation dataset as a whole, projecting predictions across the stacked trajectory to determine the point at which individual samples cross the model’s decision threshold.

4.3 Evaluation of Demographic Robustness Bias

To illustrate how the *TabMIST* framework operates as an auditing mechanism for detecting demographic robustness bias, let us consider a concrete example where the OT linear morphing process is executed independently across sub-populations. By mapping separate evaluation trajectories for each cohort, this framework allows us to directly compare geometric decision stability. Measuring



(a) Group A: Skewed distribution towards high values.

(b) Group B: Centered distribution, only slightly skewed towards high values.

Figure 4.2: *Switch Threshold* distributions computed by *TabMIST* across demographic groups on the Folktables dataset, evaluated on a sample of 5,000 instances using an Multi-Layer Perceptron (MLP) model. Group A (left) represents the minority non-white demographic cohort and exhibits a pronounced skew toward higher threshold values. Group B (right) represents the majority white demographic cohort and displays a comparatively centered distribution.

the disparities in the resulting $TabMIST_{Avg}$ and $TabMIST_{Std}$ scores serves as an illustrative case to show whether a model’s decision boundary is disproportionately fragile, unstable, or rigid for minority populations compared to a majority baseline.

As a specific example of this diagnostic capability, consider a model evaluated on two distinct groups. For the favored majority baseline (Group B), the model achieves a $TabMIST_{Avg}$ score of 53.540%, sitting close to the theoretical ideal midpoint of 50%. This indicates a well-balanced decision boundary where predictions switch following a symmetric transformation toward the target class. In contrast, the disadvantaged minority cohort (Group A) exhibits a significantly higher $TabMIST_{Avg}$ score of 62.578%. This gap highlights that instances belonging to Group A require a much larger degree of structural morphing toward the target space before the classifier changes its prediction. Meanwhile, the overall stability remains highly consistent across both groups, with a $TabMIST_{Std}$ of 24.525% for Group A and 26.586% for Group B.

While these macro-level averages expose the baseline inequity, we can extract further, insights by looking into the full *Switch Threshold* distribution of this example. Moving beyond simple summary statistics allows us to see exactly how this robustness bias behaves across different data clusters. As illustrated in Figure 4.2, the distribution curve for Group A reveals a severe skewness toward higher threshold values. This shape demonstrates that the unfavorable source class ($Y = 0$) remains heavily dominant throughout the network’s internal representation of the minority group. The model stubbornly maintains the negative prediction even under high levels of morphing towards the target space. This directional imbalance clearly demonstrates demographic robustness bias, exposing systematic skewness and an unfair model preference embedded along the decision surface.

Chapter 5

Experimental Setup

This chapter details the empirical framework designed to evaluate the proposed methodologies. The experimental design is structured around two core evaluation axes: a static baseline environment, where models are trained and evaluated on fixed ground-truth labels, and a performative feedback loop, where the model’s own past predictions progressively replace historical ground-truth labels across sequential training steps. This dual-environment structure allows us to assess not only whether fairness interventions are effective under controlled conditions but also whether they remain stable and resilient when subjected to the label corruption and distributional drift characteristic of real-world financial deployment.

- **Section 5.1: Datasets** describes the three benchmark datasets used throughout the evaluation, detailing the construction of target variables, sensitive attributes, chronological partitioning strategies, and the robust normalization scheme applied across all data environments.
- **Section 5.2: Model Architecture** specifies the feedforward neural network deployed as the base classifier across all experimental configurations.
- **Section 5.3: Evaluation Metrics** introduces the multi-dimensional metric framework used to assess model performance across predictive utility, demographic equity, and geometric boundary stability dimensions.
- **Section 5.4: Experimental ML Pipeline** formalizes the two experimental protocols, the baseline and performative setups, detailing the sequential training procedure, the label feedback mechanism, and the selection and justification of the baseline comparison methods.

5.1 Datasets

To ensure comprehensive validation across distinct data manifolds and varying degrees of class imbalance, the experimental pipeline leverages three tabular data environments: the Folktables Dataset (Ding et al. 2021), the Feedzai Bank Account Fraud Dataset (Jesus et al. 2022), and the IBM Benchmark Dataset (Altman 2021). These span a balanced census-income setting and two

severely imbalanced, real-world financial-fraud benchmarks, allowing us to study fairness and robustness behavior across both moderate and extreme class skew.

5.1.1 Folktables Dataset

The chosen configuration modifies the standard `ACSIncome` problem (Ding et al. 2021) structure by defining the following variables:

- **Target Variable (Y):** The task is framed as a binary classification problem $Y \in \{0, 1\}$ based on the engineered `Income > 50K` column. Here, $Y = 1$ denotes an individual whose annual gross income exceeds \$50,000, and $Y = 0$ represents individuals earning equal to or less than this threshold.
- **Sensitive Attribute (A):** To evaluate demographic differences, a binary protected feature `not_white` ($A \in \{0, 1\}$) is explicitly constructed by combining racial classification (`RACE`) and Hispanic origin (`HISP`) indicators. An individual is classified as baseline white ($A = 0$) exclusively if they identify as White (`RACE == 1`) and are of non-Hispanic origin (`HISP == 1`). Conversely, the sensitive demographic group ($A = 1$) isolates all individuals identifying with any minority racial classification or Hispanic/Latino ethnicity, positioning this sub-population as the target class for group-specific evaluation.

This dual-attribute extraction is necessary due to the structural taxonomy of the US Census Bureau data, which treats racial classification (`RACE`) and Hispanic origin (`HISP`) as completely separate, independent dimensions. Because the survey framework categorizes Hispanic identity as an ethnicity rather than a distinct race, a substantial portion of the population identifies racially as White while ethnically identifying as Hispanic. By intersecting both dimensions in the preprocessing pipeline, the framework isolates a rigorous baseline majority group consisting exclusively of non-Hispanic White individuals.

To capture the evolution of data distributions over time, the dataset is processed using a sequential, chronological pipeline. Data is extracted for the state of Washington across a six-year horizon spanning from 2014 to 2019 using the 1-Year estimates from the `person` survey filter. The processing pipeline treats each survey year as an independent, chronological steps (`step0` to `step5`), preserving the natural temporal progression of the population. Missing values across features are post-processed and filled using a constant value mapping to maintain a complete tabular matrix across all years.

5.1.2 Feedzai Bank Account Fraud Dataset

The Bank Account Fraud Dataset (Jesus et al. 2022) consists of a real-world, anonymized bank account application suite originally frameworked as a benchmark for evaluating non-stationary distribution shifts over time. It captures the class imbalance and evolving transactional behaviors characteristic of enterprise-grade financial fraud detection systems.

The pipeline preserves the chronological timeline of the data by processing the transaction records sequentially using the `month` tracking feature. The processing loop isolates the entries belonging to each unique month, converting them into independent chronological blocks (step₀ onwards). The engineered dataset configures target conditions and demographic groups for auditing through the following variables:

- **Target Variable (Y):** The prediction task is framed as a binary classification problem $Y \in \{0, 1\}$ based on the `fraud_bool` feature. A positive instance ($Y = 1$) denotes a verified fraudulent bank account application, while a negative instance ($Y = 0$) marks a legitimate opening request. This objective features an extreme class skew, with fraudulent instances occupying a minute minority of the overall application volume.
- **Sensitive Attribute (A):** To track age-based disparities, a binary protected attribute named `Elderly` ($A \in \{0, 1\}$) is explicitly constructed. Following the pipeline logic, consumers are classified under the sensitive demographic class ($A = 1$) if their recorded age is greater than or equal to 65 years (`customer_age` ≥ 65). Younger applicants fall under the baseline majority group ($A = 0$). Following this extraction step, the continuous `customer_age` vector is deleted to eliminate direct collinearity.

5.1.3 IBM Benchmark Dataset

The IBM Benchmark Dataset (Altman 2021) provides a synthetic simulation of industry credit card transaction logs, capturing multi-channel purchasing behaviors and complex user interactions. It is widely utilized to model financial risk under extreme class imbalances and non-stationary tracking conditions over long temporal horizons.

The engineered data space sets up distinct target parameters and demographic indicators for classification tasks:

- **Target Variable (Y):** The prediction objective is established as a binary classification problem $Y \in \{0, 1\}$ using the `Is Fraud?` attribute. Transactions flagged as fraudulent map to the positive class ($Y = 1$), while legitimate transactions map to the baseline negative class ($Y = 0$). This target reflects an acute class skew, with fraudulent instances occupying a small fraction of the overarching ledger.
- **Sensitive Attribute (A):** Demographic group partitioning is primarily centered on age-based protection criteria by creating an `Elderly` indicator variable ($A \in \{0, 1\}$). Following the boolean logic for the Account Fraud Dataset, individuals are classified under the sensitive minority group ($A = 1$) if their registered age is greater than or equal to 65 years (`Current Age` ≥ 65). Younger consumers fall into the baseline majority category ($A = 0$).

This preprocessing configuration isolates distinct behavioral properties across user age cohorts. Consumer segments over the age of 60 frequently exhibit distinct transactional signatures,

including restricted merchant categories, altered digital payment frequencies, and distinct seasonal purchasing patterns (Kemp and Erades Pérez 2023).

To prepare the raw transactional data for sequential validation, the pipeline consolidates the feature space and partitions it chronologically. First, text-based fields are standardized into numeric formats, where timestamps are converted into continuous fractional hours, currency formatting is stripped from transaction amounts, missing values are resolved via zero-imputation, and remaining categorical features are encoded using a standard label encoder. Lastly, the processed data is partitioned into consecutive three-year blocks spanning from 1997 to 2020 to establish the sequential chronological steps (step_0 to step_6) required for time-variant evaluation.

5.1.4 Robust Normalization

All datasets were normalized using a outlier-resilient preprocessing scheme that combines linear robust scaling with a non-linear bounded saturation mechanism. For each feature, the data is first centered using its median and scaled by its interquartile range ($IQR = Q_{75} - Q_{25}$), thereby preventing extreme anomalies from shifting or distorting the underlying distribution, which is highly common in heavy-tailed financial transaction manifolds. In cases of feature zero-variance where the computed IQR equals zero, the scale factor is dynamically defaulted to 1.0 to prevent division-by-zero errors. Following this linear alignment, a scaled hyperbolic tangent ($\tanh$) transformation is applied to smoothly bound extreme tail values without completely discarding their relative structural information. Mathematically, given a predefined saturation threshold $S = 3.0$, the final processed feature value is obtained via the mapping:

$$x_{\text{final}} = S \cdot \tanh\left(\frac{x - \text{median}}{S \cdot IQR}\right). \quad (5.1)$$

This normalization pipeline squashes the continuous feature space into the bounded interval $[-3.0, 3.0]$. By executing this transformation, we are mitigating gradient volatility allowing for a more stable learning process, while preserving the high-fidelity density variations.

5.2 Model Architecture

The model used in this work is a simple feedforward neural network, specifically a **MLP**. The network processes the input features through two hidden layers containing 64 and 32 units, respectively. Both hidden layers use ReLU as activation function, with an addition of dropout at a 0.2 rate in the first hidden layer. Finally, the output layer maps the features to 2 units that provide the final scores for binary classification. The network is trained for 20 epochs using a batch size of 64, a learning rate of 0.001 and the Adam optimizer.

5.3 Evaluation Metrics

The experimental validation relies on a multi-dimensional metric framework to capture the operational trade-offs between predictive capability, geometric stability, and demographic equity.

To measure the overall predictive capability of the models, we evaluate global performance using standard Accuracy and the F_1 -score. These metrics are calculated across the entire dataset at each step. Demographic equity is assessed by measuring the operational disparities between protected and baseline groups. Specifically, we track differences across groups for the **TP**, the **FP**, and **DP**. Geometric boundary stability is audited using two localized measures. First, we utilize the *TabMIST* to compute the minimal feature perturbation required to flip a model’s prediction, which quantifies local boundary thickness. Second, we measure the *Burden* of the negative outcome, which calculates the average logit distance to the decision boundary exclusively for individuals receiving an unfavorable prediction, serving as a proxy for the difficulty of achieving recourse.

Temporal Decay Analysis In addition to reporting raw values at each step, we evaluate how well each method resists performance degradation over time. We quantify this by calculating the degree of decay for each metric across the sequential timeline. This decay is computed as the direct difference between the metric’s value at the final chronological step and its value at the first step:

$$\Delta\text{Metric} = \text{Metric}_{\text{last}} - \text{Metric}_{\text{first}}. \quad (5.2)$$

This delta reveals the long-term stability and resilience of the model as it navigates performative loops or continuous data shifts.

Aggregate Model Ranking To determine the best overall method, we implement an aggregate ranking framework that evaluates models across all dimensions. For the standard performance metrics (Accuracy and F_1 -score), models are ranked based on their overall global scores. For the fairness and geometric robustness metrics, models are ranked based on the ratio or balance achieved between the demographic groups, penalizing configurations that favor one group over another. Finally, we used a weighted average to aggregate the individual ranks across all metrics into a single final macro-score for each model, providing an un-biased reflection of balanced model performance. More details about the motivation and implementation for this weighted average can be found in Chapter A.1.

5.4 Experimental ML Pipeline

This experiment evaluates the model’s resilience to feedback loops by running a sequential training and testing loop across chronological data steps. For each step i , a fresh neural network is initialized and trained on a cumulative dataset. This dataset is built by loading a fixed sample of 5,000 instances from the current period and concatenating it with samples from all previous

periods. Once trained, the model is evaluated on an unseen sample of 5,000 instances from the immediately following period ($i + 1$), measuring the metrics described previously.

The main difference between the two experimental setups lies in how the training labels (y) are assigned for the historical data chunks. In the static baseline setup, the model always trains on the original, true ground-truth labels from the dataset. In the performative setup, only the very first step ($i = 0$) uses the true ground-truth labels. For all subsequent steps, the true historical labels are replaced by the model’s own past predictions that were generated and saved during its prior evaluation phases. This simulates a live production deployment where the model’s past decisions actively overwrite the data distribution for future training cycles.

Additionally, the evaluation pipeline implements a specific post-processing safeguard during the inference phase to prevent class collapse or extreme prediction under-representation. When generating predictions on the upcoming data chunk ($i + 1$), the script checks the ratio of positive fraud predictions. If the model flags less than 4% of the instances as fraud, a fallback mechanism takes effect. It identifies the non-fraud instances with the highest network probabilities for the fraud class and manually flips their predictions to 1 until the total proportion of fraud predictions reaches exactly 4%. This threshold is consistent with operational fraud detection pipelines, where a minimum positive prediction rate is typically enforced to prevent model collapse under class imbalance.

It is important to acknowledge that the performative feedback loop implemented here operationalizes algorithmic performativity via a label feedback mechanism rather than true feature-level distribution shift. In the strict sense of Perdomo et al. (2020), performativity implies that the model’s deployment actively alters the underlying feature distribution $P(X)$. For instance, when a credit scoring model causes applicants to strategically modify their financial behavior in response to its decisions. Our pipeline captures a complementary but distinct phenomenon, where gradual corruption of the label distribution $P(Y)$ occurs as the model’s own past predictions replace ground-truth labels in subsequent training cycles. This is more precisely characterized as a label feedback loop or self-training under distribution shift, a well-documented failure mode in production ML systems where historical predictions compound over time, progressively distorting the training signal. While this constitutes a simplification of full performativity, the feature distributions across chronological data steps reflect genuine temporal drift in the underlying population rather than model-induced behavioral change. Therefore, this implementation captures the core mechanism of fairness decay under deployment, this being the erosion of label quality through model feedback and the amplification of existing demographic disparities as the classifier’s past errors are treated as ground truth. This operationalization is therefore defended as a tractable and empirically grounded proxy for studying performative fairness degradation in tabular financial domains, where true feature-level causal intervention would require controlled field experiments beyond the scope of this work.

Methods

To rigorously assess the performance of the proposed *FairMorph* framework, we compare two implementations of our framework, $\text{FairMorph}_{1\text{epoch}}$ where we use our augmentation pipeline every epoch of the training, and $\text{FairMorph}_{5\text{epoch}}$ where we only employ the augmentation every 5th epoch of training. The other baseline methods consist of a diverse set of state-of-the-art in-processing and data augmentation baselines. The selected baselines reflect a deliberate focus on in-processing comparators rather than an exhaustive benchmark across all fairness intervention paradigms. Pre-processing and post-processing approaches were excluded on the grounds that neither operates directly on the decision boundary during training, which is precisely the mechanism *FairMorph* targets, comparing against them would conflate the evaluation with a broader benchmark study rather than a focused assessment of boundary-targeted augmentation. Within the in-processing family, *FaiR-N* was selected as the closest prior work explicitly targeting the fairness-robustness intersection through a geometric boundary-based objective, while *Fair Mixup*, specifically the **EO** and **DP** variants, serve as natural interpolation-based augmentation baselines, isolating the specific contribution of **OT** matching and adversarial boundary targeting over naive demographic mixing.

FaiR-N_{1,0} Proposed by Sharma et al. (2021), the *FaiR-N* baseline alters the optimization landscape by incorporating logit-approximated boundary distances into the objective function. In this configuration, the hyperparameters are skewed strictly toward demographic equity ($\lambda_F = 1.0, \lambda_R = 0.0$), forcing the network to minimize exclusively the disparity in the average burden of recourse between protected demographic groups. This variant isolates the impacts of optimizing purely for recourse-based fairness without explicit robustness constraints.

FaiR-N_{0.5,0.5} This configuration of the *FaiR-N* architecture balances the objective function equally between the dual regularizers ($\lambda_F = 0.5, \lambda_R = 0.5$). The model simultaneously regularizes the network to equalize the average distance to the decision boundary for disfavored instances while maximizing the global margin (I_{robust}) across all training samples. It serves as a baseline to evaluate the architectural trade-offs between global adversarial robustness, accuracy preservation, and demographic equity.

Fair Mixup DP Proposed by Chuang and Mroueh (2021), this baseline implements a data augmentation strategy that enforces **DP** by regularizing the classifier on continuous paths of interpolated samples connecting the sensitive groups P_0 and P_1 . Utilizing input mixup as a linear interpolator $T(x_0, x_1, t) = tx_0 + (1-t)x_1$, the framework optimizes a closed-form smoothness regularizer $R_{\text{mixup}}^{\text{DP}}(f)$ alongside the standard classification loss. This penalty evaluates the expected inner product between the model’s Jacobian on the augmented samples and the geometric difference vector $x_0 - x_1$. This explicitly limits predictive fluctuations to ensure the function remains invariant as sensitive attributes change across the input manifold. Implementation-wise, a single interpolation weight t is sampled per batch to maintain computational efficiency.

Fair Mixup EO This variant extends the interpolation framework to satisfy equalized odds constraints by conditioning the generative paths strictly on the ground-truth label $y \in \{0, 1\}$. Rather than mixing globally across entire demographic cohorts, the regularizer $R_{\text{mixup}}^{\text{EO}}(f)$ computes linear interpolations conditioned between the group-specific conditional distributions P_0^y and P_1^y . By penalizing the directional derivative of the model’s expected predictions exclusively within these label-matched blocks, this baseline targets joint stabilization of **TP** and **FPrates** across the intermediate spaces. This forces the model to homogenize its error profiles across groups, enhancing the generalization of **EO** constraints at evaluation time.

Vanilla Model Lastly, the *Vanilla* baseline represents standard empirical risk minimization trained exclusively using the classical cross-entropy loss function ($\mathcal{L}_{\text{cross}}$) without any fairness interventions or robustness regularization terms. It serves as the primary control baseline to establish maximum unconstrained predictive capacity, mapping out the native accuracy, systemic fairness disparities, and geometric vulnerabilities inherent to the raw data distribution.

Chapter 6

Results and Discussion

This chapter presents and discusses the empirical findings obtained across the three benchmark datasets under both static and performative evaluation environments. Results are organized around the four research questions introduced in Chapter 1, progressing from the characterization of performative fairness decay through the multi-objective comparison of *FairMorph* against state-of-the-art baselines to the diagnostic utility of the *TabMIST* auditing framework and its capacity to expose pathological optimization shortcuts in severely imbalanced fraud domains.

- **Section 6.1: RQ1** investigates whether group-level equity criteria satisfied during static training persist when models are deployed in active feedback loops where past predictions overwrite subsequent training labels.
- **Section 6.2: RQ2** evaluates the multi-objective rank stability and decision-boundary resilience of *FairMorph* relative to in-processing regularizers and interpolation-based augmentation baselines under performative stress.
- **Section 6.3: RQ3** assesses whether probing decision boundaries via continuous sample-level morphing uncovers topological anomalies that aggregate macro-level fairness metrics fail to detect.
- **Section 6.4: Limitations** discusses the structural and methodological constraints of the proposed framework, identifying open challenges and boundary conditions under which the reported findings may not generalize.

6.1 RQ1: Do fairness metrics decay over performative loops in tabular distributions?

Our analysis confirms that fairness measured by standard metrics is unreliable under performative loops, it either decays or improves spuriously via class collapse. While static evaluations often suggest that bias mitigation strategies are effective, these metrics frequently mask underlying structural degradation that occurs as the model interacts with its own predictions.

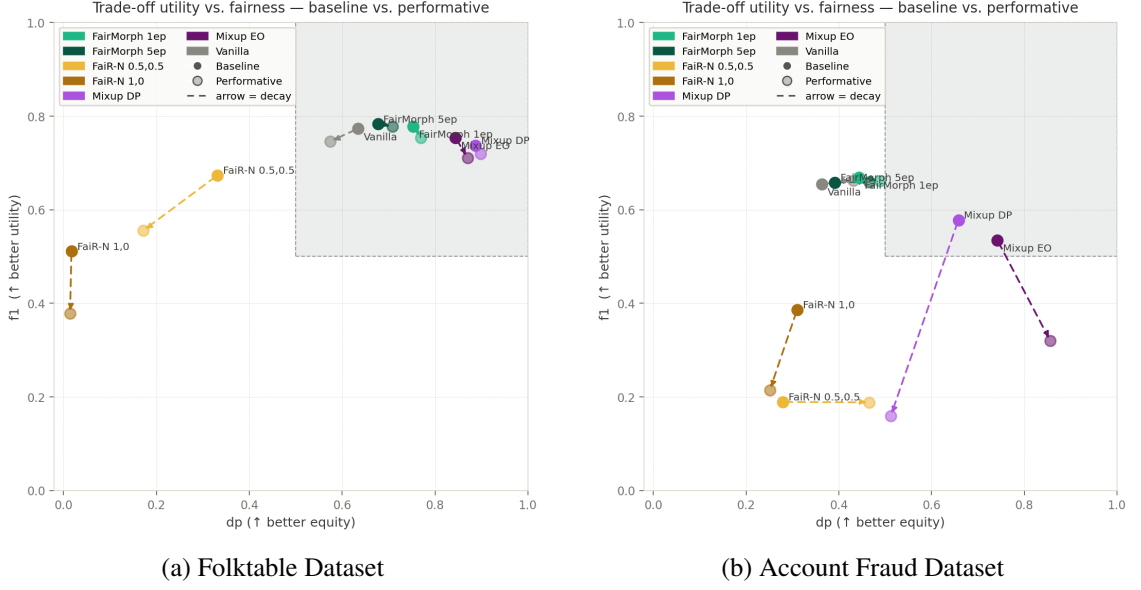


Figure 6.1: Trade-off between predictive utility (F_1 -score) and DP. Each point represents the performance of a specific model configuration. Darker, solid circles denote performance in the static baseline environment, while lighter, hollow circles denote performance under the performative feedback loop. Dashed lines (arrows) indicate the directional decay of model performance over the performative horizon.

6.1.1 Label Feedback Loop Decay

This erosion is visualized in Figure 6.1, where the dashed decay arrows illustrate the transition from static baseline performance to performative deployment. Across all tested paradigms, we observe a consistent leftward shift in DP, indicating that performative loops actively drive models away from ideal boundary equity and amplify systemic bias. A closer inspection of Figure 6.1a reveals that while no single method achieves a global improvement on both objectives simultaneously, different paradigms exhibit distinct behavioral trajectories. The *Vanilla* model and several unconstrained baselines suffer from simultaneous degradation, while the *FairMorph* and *Mixup* variants demonstrate greater robustness, successfully navigating the performative landscape by maintaining baseline F_1 levels while simultaneously achieving a slight increase in DP ratio. This suggests that while these models are not immune to the performative loop, they possess a structural plasticity that allows them to re-center their decision boundaries, often trading marginal utility gains to preserve or improve group equity.

Even when explicit regularization is applied, performative feedback triggers structural instability in many cases. In the severely imbalanced Account Fraud setting, the *FaiR-N* variants demonstrate that satisfying fairness constraints in a static environment does not guarantee long-term stability. As shown by the downward trajectories in Figure 6.1b, methods like *Fair Mixup* and *FaiR-N* often degrade heavily in both utility and equity simultaneously, suggesting that current mitigation strategies are highly sensitive to the temporal drift inherent in feedback loops.

6.1.2 Decay Over Time

Beyond the static-to-performative transition, a second and equally important axis of degradation emerges when examining how metrics evolve within each experiment across sequential training steps. Tables B.13 through B.18 report the temporal delta $\Delta = \text{Metric}_{\text{last}} - \text{Metric}_{\text{first}}$ for each method across all datasets and environments, capturing how much each metric drifts between the first and final training step. Crucially, because all three datasets are partitioned into chronological slices reflecting genuine temporal progression (census survey years for Folktables, three-year transaction blocks for IBM, and monthly application windows for the Feedzai dataset) this within-experiment decay is not merely a consequence of label feedback corruption. It also captures the compounding effect of real distributional shift in the underlying population, where the feature distribution $P(X)$ evolves naturally over time independently of the model’s predictions.

On the Folktables dataset under the performative loop (Table B.14), the unregularized *Vanilla* baseline undergoes consistent multi-objective erosion ($\Delta\text{Acc} = -0.065$, $\Delta F_1 = -0.144$, $\Delta\text{DP} = -0.131$), confirming that naive feedback loops inherently amplify systemic bias over time. Explicit regularization does not resolve this: *FaiR-N*_(0.5,0.5) suffers a catastrophic utility collapse ($\Delta F_1 = -0.461$), while *Mixup DP* and *Mixup EO* experience substantial fairness erosion ($\Delta F_1 = -0.205$ and -0.258 respectively). In contrast, *FairMorph* variants resist this temporal drift, with *FairMorph*_{5epoch} restricting utility decay to $\Delta F_1 = -0.070$ and *FairMorph*_{1epoch} actively consolidating equity over time ($\Delta\text{DP} = +0.237$).

Critically, the IBM Transaction Fraud tables (Tables B.17 through B.18) illustrate the spurious improvement phenomenon flagged above: *FaiR-N*_(0.5,0.5) and *Mixup EO* post positive accuracy and DP deltas under the performative loop, which superficially suggest improvement. However, their *TabMIST*_{Avg} deltas collapse to zero entirely, confirming that these apparent gains are artifacts of single-class default rather than genuine learning.

Taken together, both decay axes, the transition from static to performative deployment and the within-experiment temporal drift, demonstrate that fairness is not a static property of a trained classifier but a dynamic state vulnerable to feedback-induced corruption. Standard mitigation methods fail to account for this temporal instability across the full multi-objective surface, with secondary dimensions such as EO disparities and structural Recourse Burden scores undergoing widespread decay alongside primary utility and fairness metrics (Tables B.13 through B.18). These findings confirm that fairness-aware models require explicit, boundary-focused reinforcement to resist the systemic erosive effects of long-term deployment.

6.2 RQ2: How does the proposed *FairMorph* framework compare against existing state-of-the-art mitigation strategies under performative stress?

Building directly upon the performative decay insights established in RQ1, where standard architectures demonstrated systemic vulnerabilities under feedback stress, this section evaluates

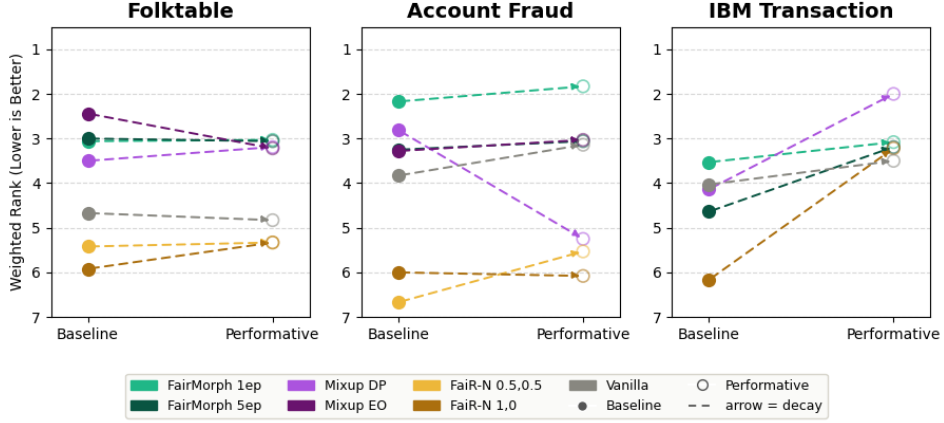


Figure 6.2: Multi-objective Weighted Rank stability across environments. The subplots trace the evolution of the Weighted Rank from the static baseline to the dynamic performative environment across the Folktable, Account Fraud, and IBM Transaction Fraud datasets.

how the proposed *FairMorph* framework performs relative to contemporary state-of-the-art in-processing and data-augmentation baselines (Perdomo et al. 2020). To achieve a reliable, comprehensive comparison across the competing demands of predictive utility, group equity, and decision surface robustness, we analyze the consolidated multi-objective Weighted Rank profiles across both static and performative execution environments. This comprehensive comparison is visually synthesized in Figure 6.2, which traces the ranking trajectories of all competing methods across the three benchmark domains.

The primary operational advantage of *FairMorph* is not a dramatic improvement in any single static fairness metric. Rather, its distinguishing property is resilience: it preserves a stable, well-defined decision boundary across the transition from static training to performative deployment, where competing methods either lose rank or break down entirely. As illustrated in the Folktable subplot of Figure 6.2, the *FairMorph* variants maintain an almost flat Weighted-Rank profile, securing the top rank (Rank = 1.0) for both Accuracy and F1-score across baseline and performative loops (Tables B.7 through B.8), while their fairness ratios neither collapse nor inflate spuriously.

The contrasting behavioral trajectories of traditional methods highlight this resilience; while *Mixup EO* leads the static baseline with an optimal Weighted Rank of 2.44, it undergoes a significant multi-objective degradation under performative conditions, dropping down to a Weighted Rank of 3.22, which is reflected by its downward slope in Figure 6.2. Conversely, both *FairMorph*_{1epoch} and *FairMorph*_{5epoch} effectively anchor the performative environment, preserving highly stable consolidated positions at Weighted Ranks of 3.03 and 3.06, respectively.

Imbalanced and Dynamic Fraud Domains

This multi-objective resilience becomes explicitly pronounced when transitioning to heavily skewed distributions, which are typical of real-world risk settings. In the Account Fraud scenario (Fig-

ure 6.2, center), *FairMorph*_{1epoch} demonstrates an upward-sloping trajectory, indicating that its relative performance advantage actually expands under performative stress. In the static baseline scenario (Table B.9 and B.10), *FairMorph*_{1epoch} dominates the evaluation pool, claiming the top spot (Rank = 1.0) for Accuracy, F_1 -score, and $TabMIST_{Avg}$ (M_{Mean}), resulting in a superior overall Weighted Rank of 2.17.

Once the performative feedback loop is active (Table B.10), rather than decaying alongside baselines like *Mixup DP*, which suffers rank collapse from 2.81 down to 5.25, *FairMorph*_{1epoch} extends its competitive advantage further, achieving the highest Weighted Rank of 1.83. In this deployment environment, it secures the top rank (Rank = 1.0) for **TP** rate alignment, M_{Mean} , and M_{Std} , empirically confirming that it constructs the most spatially uniform and fair decision surface under severe distribution shifts.

As detailed in the tabular rankings of the highly skewed IBM Transaction Fraud dataset (Table B.11 and B.12) and visualized in the final subplot of Figure 6.2, traditional in-processing strategies such as *FaiR-N*_(0.5,0.5) and data-smoothing techniques like *Mixup EO* break down entirely under performative stress, collapsing into single-class predictors. In sharp contrast, by utilizing active batch-level **OT** matching to pair vulnerable instances with their closest counter-demographic counterparts, *FairMorph* preserves its active classification boundary.

Operating inline during network training, it consistently places in the top tier for micro-boundary consistency, securing an $TabMIST_{Avg}$ rank of 2.0 for the 1-epoch variant and 1.0 for the 5-epoch variant under the performative IBM loop. Ultimately, these consolidated rankings and visual trajectories verify that *FairMorph* establishes a dependable equilibrium between all metrics.

6.3 RQ3: Does the introduction of the *TabMIST* framework provide tangible benefits to empirical model auditing?

Traditional algorithmic fairness auditing relies almost exclusively on macro-level statistical dependencies calculated over aggregate populations. While metrics such as **DP** and **EO** are valuable tracking tools, they operate as black boxes, evaluating outcomes rather than the underlying model topology. This section demonstrates that the proposed *TabMIST* framework introduces a high-resolution diagnostic layer by shifting the auditing paradigm from global population ratios to micro-level geometric decision boundary probing.

Quantifying Boundary Drift and Spatial Neutrality

An empirical benefit of the *TabMIST* diagnostic suite is its capacity to unmask optimization shortcuts that deceptively skew traditional metrics. This phenomenon is starkly illustrated in the skewed IBM Transaction Fraud dataset under performative stress, captured in the comparative subplots of Figure 6.3. When auditing via standard macro-level metrics alone, as shown in Figure 6.3a, regularizers like *FaiR-N* and data-smoothing techniques like *Mixup EO* appear to preserve or even enhance their fairness profile under performative deployment. These configurations are pushed

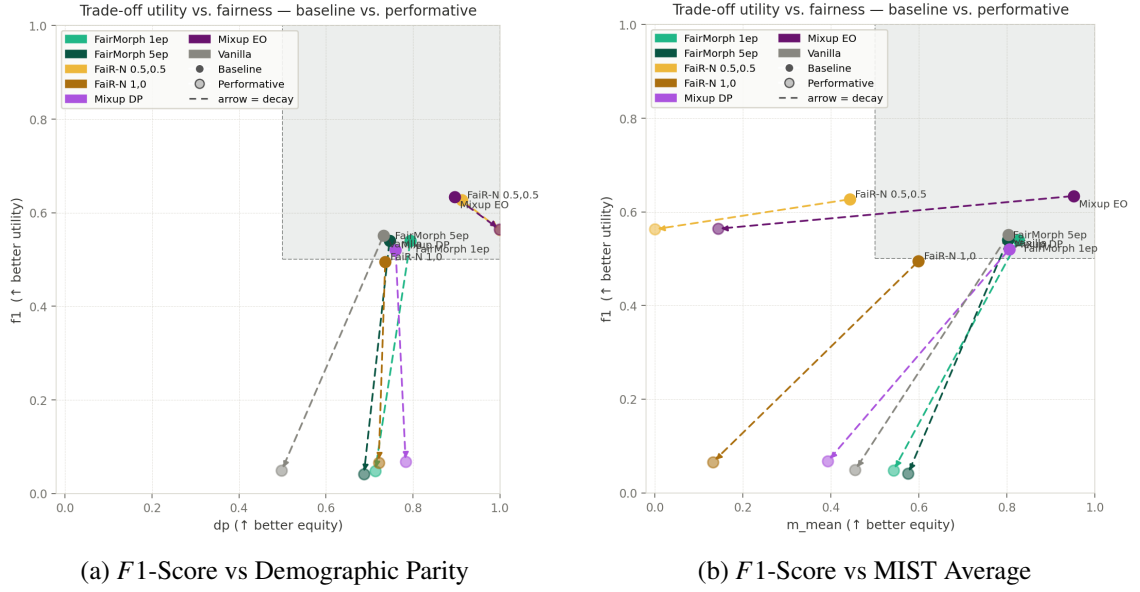


Figure 6.3: Trade-off between predictive utility (F_1 -score) and fairness for IBM Transaction Fraud Data. Each point represents the performance of a specific model configuration. Darker, solid circles denote performance in the static baseline environment, while lighter, hollow circles denote performance under the performative feedback loop. Dashed lines (arrows) indicate the directional decay of model performance over the performative horizon.

to the far right of the plot, achieving high **DP** values approaching 1.000. Intuitively, this outward trajectory is deeply misleading; it falsely signals to an auditor that the deployment loop has consolidated group equity while maintaining a stable utility trade-off.

Cross-referencing this behavior with the continuous spatial diagnostics of the *TabMIST* framework immediately dismantles this algorithmic illusion. As shown in Figure 6.3b, tracking the identical performative transition through the lens of the $TabMIST_{Avg}$ (m_mean) reveals a straight slide to the far left of the manifold, approaching a value of zero. The mechanistic reason for this collapse is precise: *TabMIST* requires valid source-to-target prediction transitions along each interpolation trajectory, meaning it can only compute a *Switch Threshold* for a given pair when the classifier actually changes its prediction somewhere along the morphing path. A model that has defaulted to predicting the majority class universally produces no such transitions. The metric does not return a misleading value; it returns silence, which is itself an unambiguous diagnostic signal. Consequently, $TabMIST_{Avg}$ collapses to zero not as an arbitrary floor, but as a direct mechanistic consequence of decision boundary erasure.

The high **DP** scores achieved in Figure 6.3a are therefore not artifacts of an equitable, robust decision boundary; rather, they are structural pathologies where the classifier has satisfied global demographic constraints by completely defaulting into a trivial majority-class assigner. Because the model uniformly predicts the minority class across all instances, group-level statistical disparities mathematically vanish, hiding a catastrophic breakdown of actual model utility. Crucially, this is the failure mode that aggregate fairness metrics cannot detect, a model with perfect **DP** and

zero utility is indistinguishable from a genuinely fair model when evaluated through the lens of population-level outcome ratios alone.

Ultimately, these evaluations confirm that macro-level fairness constraints are highly vulnerable to being exploited by *DNNs* seeking the path of least resistance during optimization. By forcing the evaluation suite to measure the continuous spatial characteristics of local instances rather than global outcomes, *TabMIST* serves as an ungameable spatial auditor. It provides engineers with high-fidelity diagnostics to guarantee that a model’s fairness profile reflects genuine structural boundary equity rather than the artifact of single-class default, which provides a safeguard in high-stakes financial deployment where the consequences of mistaking model collapse for model fairness are severe.

6.4 Limitations

While the *FairMorph* framework demonstrates superior rank stability and boundary resilience under performative feedback loops, several inherent limitations must be considered when interpreting these results.

First, the core mechanism of *FairMorph* relies on *OT* to identify counter-demographic counterparts, solving a minimum-cost assignment between minority and majority instances in the feature space. This implicitly assumes that geometrically proximate instances across demographic groups should share the same ground-truth label, which treats proximity as a proxy for class similarity. While this assumption is grounded in the broader manifold hypothesis and is empirically supported by our results across tabular fraud domains, it remains a heuristic rather than a guarantee.

Second, although the augmentation strategy is constrained to injecting at most one adversarial sample per source instance per epoch, these generated instances are linear interpolations between empirical data points and therefore lie within the convex hull of the training data rather than on the true data manifold. Real-world tabular distributions, particularly in financial fraud detection, are inherently non-convex, characterized by discrete feature dependencies and low-density boundary regions that linear interpolation does not faithfully reconstruct. As a result, the synthetic samples may not perfectly reflect the statistical properties of the underlying data generating process.

Third, unlike traditional in-processing regularizers such as *FaiR-N*, which optimize directly against explicit statistical parity constraints through fixed penalty terms, *FairMorph* operates as a dynamic adversarial augmentation framework. By continuously probing the model’s evolving decision boundary and injecting targeted boundary-crossing synthetic instances, it drives the classifier toward demographic invariance through geometric pressure rather than explicit constraint satisfaction. As a consequence, *FairMorph* does not provide mathematical guarantees that the model will satisfy a hard statistical fairness threshold at any given training step, a property that threshold-based regularizers nominally offer. For deployment contexts subject to strict regulatory compliance requirements, where auditors may demand provable satisfaction of specific parity criteria, this absence of explicit guarantees may present a practical obstacle. That said, our empirical results suggest that rigid parity constraints are themselves vulnerable to pathological satisfaction

Table 6.1: Computational overhead of the *FairMorph* augmentation step across datasets, measured on a fixed sample size of 5,000 instances per training step. Morphing time is reported as the wall-clock time spent exclusively within the `fit_morph_linear` call. Overhead ratio is computed as the ratio of total augmented training time to total vanilla training time.

Dataset	Method	Epoch 4 (s)	Epoch 9 (s)	Epoch 14 (s)	Epoch 19 (s)	Total Morph Time (s)
Folktables	Augmented	0.1417	0.1373	0.1390	0.1361	0.5541
Account Fraud	Augmented	0.0021	0.0023	0.0022	0.0014	0.0080
IBM Transaction	Augmented	0.6042	0.6173	0.6363	0.6359	2.4937

Dataset	Total Augmented Time (s)	Total Vanilla Time (s)	Overhead Ratio
Folktables	12.756	10.119	1.26×
Account Fraud	12.046	10.126	1.19×
IBM Transaction	13.757	8.122	1.69×

through majority-class collapse under performative stress, raising questions about the practical value of such guarantees in dynamic deployment settings.

Finally, to empirically characterize the computational overhead introduced by the batch-level **OT** matching step, we measured the wall-clock morphing time across all three datasets on a fixed sample size of 5,000 instances per training step. Results are summarized in Table 6.1. The total morphing time across the full training horizon represented a mean overhead ratio of 1.26×, 1.00×, and 1.69× relative to the Vanilla baseline for the Folktables, Account Fraud, and IBM Transaction Fraud datasets respectively. The negligible overhead on the Account Fraud dataset reflects the extreme class imbalance inherent to that domain, where the minority misclassification set is frequently too small to trigger the augmentation step, resulting in near-zero morphing cost. Conversely, the higher overhead on the IBM dataset reflects the larger source and target sets constructed at each augmentation epoch, which increase the number of **OT** pairs the Hungarian algorithm must resolve. Notably, the per-augmentation-epoch morphing time remains stable across training epochs within each dataset, with low variance across the sampled checkpoints (epochs 4, 9, 14, and 19), suggesting the overhead is predictable and does not compound over the performative horizon. While the 1.69× overhead on IBM represents a meaningful increase in training time, it remains within acceptable bounds for an offline retraining pipeline typical of financial fraud detection systems, where model updates occur periodically rather than in real time.

Chapter 7

Conclusions and Future Work

This final chapter synthesizes the core findings, methodological advancements, and empirical insights established throughout this thesis. By unifying the fields of algorithmic fairness and model robustness within highly dynamic and imbalanced tabular domains, this research addresses structural vulnerabilities in contemporary **RAI** deployment. This chapter reviews our explicit findings relative to our guiding research questions, formalizes the primary contributions of this work, addresses existing architectural limitations, and outlines prospective avenues for future investigation.

7.1 Definitive Synthesis of Research Questions

RQ1: Performative Metric Decay Our evaluations confirm that fairness metrics are unreliable under performative feedback loops. Models satisfying group fairness constraints on ground-truth labels frequently degrade once their own past predictions corrupt the training signal. The Vanilla baseline experiences simultaneous degradation across utility ($\Delta F1 = -0.144$) and equity ($\Delta DP = -0.131$) on Folktables, confirming that naive feedback loops amplify systemic bias. More troublingly, several methods (*FaiR-N*, *Mixup EO*) exhibit spurious metric improvements under performative stress on the IBM dataset, with positive accuracy and **DP** deltas masking complete decision boundary collapse. Standard auditing provides no protection against this failure mode. *FairMorph* avoids this collapse by enforcing geometric pressure toward demographic invariance rather than explicit statistical constraint satisfaction, maintaining active multi-class learning throughout deployment.

RQ2: FairMorph Multi-Objective Efficacy *FairMorph* distinguishes itself not through dramatic single-metric improvements but through maintaining a stable Weighted Rank performance across the transition from static to performative environments, where competing methods degrade significantly. On Account Fraud, *FairMorph*_{1epoch} achieves Weighted Rank 1.83 under performative stress while Mixup **DP** collapses from 2.81 to 5.25. Critically, on the most severely imbalanced IBM dataset, traditional methods collapse into single-class predictors while *FairMorph* preserves its decision boundary.

RQ3: *TabMIST* Diagnostic Utility Macro-level fairness metrics are insufficient for detecting optimization pathologies in production. On IBM Transaction Fraud, *FaiR-N* and Mixup **EO** appear to preserve fairness ($DP \approx 1.000$) when audited through standard metrics, yet *TabMIST*_{Avg} slides to zero, revealing decision boundary erasure. *TabMIST* cannot be fooled by majority-class default, our method produces silence, not favorable scores. *TabMIST* serves as an ungameable spatial auditor, providing assurance that fairness profiles reflect structural equity rather than model collapse.

7.2 Summary of Primary Contributions

This thesis delivers three primary contributions to the engineering and ethical deployment of machine learning in high-stakes financial risk assessment:

1. **The FairMorph Framework:** An inline, batch-level data augmentation methodology inspired by the straight-line trajectory principles of **OT-CFM**. Rather than training heavy parameterized neural vector fields to model global data flows, *FairMorph* operationalizes optimal transport geometry directly within active training batches, identifying cross-demographic neighbor pairings and generating targeted, boundary-crossing synthetic instances. The framework demonstrates superior multi-objective rank stability under performative stress, particularly in severely imbalanced fraud domains where existing methods completely fail.
2. **The *TabMIST* Auditing Paradigm:** A geometric boundary diagnostic lens capable of identifying spatial boundary erasure and decision fragility that aggregate fairness metrics systematically conceal. By measuring the continuous interpolation distance required to flip classifier predictions along inter-class morphing trajectories, *TabMIST* provides an ungameable spatial auditor that cannot be exploited through trivial majority-class default, offering high-fidelity assurance that a model’s fairness profile reflects genuine structural equity.
3. **A Dynamic Performative Benchmark and Empirical Characterization:** A comprehensive evaluation framework documenting classifier degradation across static baselines and performative loops in three distinct financial fraud datasets, ranging from balanced census-income distributions to severely imbalanced transaction environments. This empirical work establishes baseline understanding of how fairness-robustness trade-offs evolve under realistic deployment pressures, quantifies the vulnerability of existing methods to label feedback corruption and temporal distribution drift, and provides concrete evidence that fairness constraints can be satisfied through pathological optimization shortcuts in extreme class-imbalance settings.
4. **Published and Ongoing Extensions to the *tsMIST* Framework:** The morphing-based sensitivity analysis paradigm underlying *TabMIST* was first introduced in the time-series domain as *tsMIST* (Brito et al. 2025), published at the International Conference on Discovery Science 2025. Further research developed in this thesis was subsequently extended and

submitted for publication, with an acceptance received in June 2026. This extended work is currently under active revision to incorporate reviewer feedback.

Systemic Limitations Several limitations merit acknowledgment. First, *FairMorph* relies on geometric proximity as a proxy for class similarity, which may not hold in domains where feature geometry misaligns with class structure. Second, synthetic samples lie within the convex hull of training data rather than on the true manifold, potentially introducing distribution shift in non-convex tabular domains. Third, *FairMorph* provides no hard statistical fairness guarantees, relying instead on geometric pressure toward demographic invariance. For regulatory contexts demanding provable satisfaction of specific parity criteria, this absence may present practical obstacles. Fourth, batch-level OT introduces computational overhead of $1.26\times$ to $1.69\times$ relative to baselines, which may be prohibitive in real-time scenarios. Finally, the performative feedback loops operationalize label corruption rather than true feature-level performativity. While this simplification remains tractable and empirically grounded, it represents a partial operationalization of algorithmic performativity as formally defined in prior work.

7.3 Future Work

Several concrete directions warrant investigation. First, extending both *FairMorph* and *TabMIST* to image domains would test whether optimal transport matching and boundary morphing principles generalize beyond tabular data. Second, extending to multi-sensitive attribute scenarios would address intersectional fairness, where demographic robustness bias manifests across multiple protected dimensions simultaneously. Third, deeper investigation into how augmentation across different manifold density regions affects the underlying data distribution would be valuable. Our initial sampling topology study (Section 3.3) revealed trade-offs between data fidelity and geometric regularization; future work should systematically characterize whether targeting specific density areas provides task-dependent benefits beyond demographic fairness. Fourth, adapting the performativity experiment toward more realistic feedback mechanisms would strengthen ecological validity. The current label corruption pipeline represents a simplification; operationalizing true feature-level performativity where model predictions induce behavioral responses from affected populations would better capture real-world deployment dynamics. Finally, broader empirical validation across diverse datasets without performative stress would establish whether *FairMorph* and *TabMIST* provide genuine benefits in non-feedback settings or whether their advantages are specific to performative environments. This would clarify the true generalizability of both methods beyond the high-stakes fraud detection context that motivated this work.

7.4 Concluding Remarks

Algorithmic decision-making systems actively shape access to financial services, credit, and opportunity for millions globally. As these systems become embedded in feedback loops where pre-

dictions influence subsequent data distributions, the classical paradigm of static training becomes fundamentally inadequate. This thesis demonstrates that fairness is not a stable engineered property but a dynamic state that erodes under performative pressure. Traditional fairness constraints can be satisfied through pathological shortcuts (like single-class default) that standard metrics fail to expose, particularly in severely imbalanced domains. *FairMorph* and *TabMIST* address this challenge by shifting from static constraint satisfaction toward continuous geometric reinforcement and from population-level metrics toward micro-level boundary interrogation. While these contributions represent meaningful progress, **RAI** ultimately requires integrated approaches unifying fairness, robustness, transparency, and accountability throughout the model lifecycle. As algorithmic systems become increasingly consequential, balancing utility, equity, and long-term performative resilience is no longer optional but a fundamental necessity. *FairMorph* provides a dependable methodological step toward systems designed with affected communities at their center.

References

- Abdallah, Aisha, Mohd Aizaini Maarof, and Anazida Zainal. 2016. "Fraud detection system: A survey." *Journal of Network and Computer Applications* 68:90–113. <https://www.sciencedirect.com/science/article/pii/S1084804516300571>.
- Abernethy, Jacob D, Elad Hazan, and Alexander Rakhlin. 2009. "Competing in the dark: An efficient algorithm for bandit linear optimization." In *Conference on Learning Theory*. 110. <https://repository.upenn.edu/entities/publication/1ec8ecd4-3eb9-4b92-9464-5acaa5a8bf18>.
- Ahmad, Muhammad Aurangzeb, Carly Eckert, and Ankur Teredesai. 2018. "Interpretable machine learning in healthcare." In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, 559–560. <https://dl.acm.org/doi/10.1145/3233547.3233667>.
- Albergo, Michael S, and Eric Vanden-Eijnden. 2022. "Building normalizing flows with stochastic interpolants." *arXiv preprint arXiv:2209.15571*, <http://arxiv.org/abs/2209.15571>.
- Altman, Erik. 2021. "Synthesizing credit card transactions." In *Proceedings of the Second ACM International Conference on AI in Finance*, 1–9. <https://dl.acm.org/doi/abs/10.1145/3490354.3494378>.
- Balendran, A, C Beji, F Bouvier, O Khalifa, T Evgeniou, P Ravaud, and R Porcher. *A scoping review of robustness concepts for machine learning in healthcare*. *NPJ Digit. Med.* 8 (1)(2025). <https://www.nature.com/articles/s41746-024-01420-1>.
- Barocas, Solon, and Andrew D Selbst. 2016. "Big data's disparate impact." *Calif. L. Rev.* 104:671. <https://www.jstor.org/stable/24758720>.
- Bartlett, Peter L, Dylan J Foster, and Matus J Telgarsky. 2017. "Spectrally-normalized margin bounds for neural networks." *Advances in neural information processing systems* 30. https://proceedings.neurips.cc/paper_files/paper/2017/file/b22b257ad0519d4500539da3c8bcf4dd-Paper.pdf.
- Beede, Emma, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamvi-boonsuk, and Laura M Vardoulakis. 2020. "A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy." In *Proceedings of the 2020 CHI conference on human factors in computing systems*, 1–12. <https://dl.acm.org/doi/abs/10.1145/3313831.3376718>.
- Blanchet, Jose, Yang Kang, and Karthyek Murthy. 2019. "Robust Wasserstein profile inference and applications to machine learning." *Journal of Applied Probability* 56 (3): 830–857. <https://www.cambridge.org/core/journals/journal-of-applied-probability/article/abs/robust-wasserstein-profile-inference-and-applications-to-machine-learning/4024D05DE4681E67334E45D039295527>.

- Bolukbasi, Tolga, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. “Man is to computer programmer as woman is to homemaker? debiasing word embeddings.” *Advances in neural information processing systems* 29. https://proceedings.neurips.cc/paper_files/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html.
- Brito, Antónia, Moisés Santos, Duarte Folgado, and Carlos Soares. 2025. “tsMIST: Model Sensitivity Analysis with Time Series Morphing.” In *International Conference on Discovery Science*, 270–285. Springer. https://doi.org/10.1007/978-3-032-05461-6_18.
- Calders, Toon, Faisal Kamiran, and Mykola Pechenizkiy. 2009. “Building classifiers with independence constraints.” In *2009 IEEE international conference on data mining workshops*, 13–18. IEEE. <https://ieeexplore.ieee.org/abstract/document/5360534>.
- Chakraborty, Anirban, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. 2021. “A survey on adversarial attacks and defences.” *CAAI Transactions on Intelligence Technology* 6 (1): 25–45. <https://ietresearch.onlinelibrary.wiley.com/doi/full/10.1049/cit2.12028>.
- Chen, Ricky TQ, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. 2018. “Neural ordinary differential equations.” *Advances in neural information processing systems* 31. https://proceedings.neurips.cc/paper_files/paper/2018/hash/69386f6bb1dfed68692a24c8686939b9-Abstract.html.
- Chouldechova, Alexandra. 2017. “Fair prediction with disparate impact: A study of bias in recidivism prediction instruments.” *Big data* 5 (2): 153–163. <http://arxiv.org/abs/1610.07524>.
- Chuang, Ching-Yao, and Youssef Mroueh. 2021. “Fair mixup: Fairness via interpolation.” *arXiv preprint arXiv:2103.06503*, <http://arxiv.org/abs/2103.06503>.
- Correia, André, Carlos Soares, and Alípio Jorge. 2019. “Dataset Morphing to Analyze the Performance of Collaborative Filtering.” In *Discovery Science: 22nd International Conference, DS 2019, Split, Croatia, October 28–30, 2019, Proceedings*, 29–39. Split, Croatia: Springer-Verlag. ISBN: 978-3-030-33777-3. https://doi.org/10.1007/978-3-030-33778-0_3. https://doi.org/10.1007/978-3-030-33778-0_3.
- Cotter, Andrew, Maya Gupta, Heinrich Jiang, Nathan Srebro, Karthik Sridharan, Serena Wang, Blake Woodworth, and Seungil You. 2018. “Training fairness-constrained classifiers to generalize.” In *ICML 2018 workshop: “fairness, accountability, and transparency in machine learning (FAT/ML)*. https://www.fatml.org/media/documents/training_fairness_constrained_classifiers_to_generalize.pdf.
- Dignum, Virginia. 2019. *Responsible artificial intelligence: how to develop and use AI in a responsible way*. Vol. 2156. Springer. <http://link.springer.com/10.1007/978-3-030-30371-6>.
- Ding, Frances, Moritz Hardt, John Miller, and Ludwig Schmidt. 2021. “Retiring Adult: New Datasets for Fair Machine Learning.” *Advances in Neural Information Processing Systems* 34. https://proceedings.neurips.cc/paper_files/paper/2021/hash/32e54441e6382a7fbacbbbf3c450059-Abstract.html.
- Dressel, Julia, and Hany Farid. 2018. “The accuracy, fairness, and limits of predicting recidivism.” *Science advances* 4 (1): eaao5580. <https://www.science.org/doi/full/10.1126/sciadv.aao5580>.

- Du, Wei, and Xintao Wu. 2021. “Fair and robust classification under sample selection bias.” In *Proceedings of the 30th acm international conference on information & knowledge management*, 2999–3003. <https://dl.acm.org/doi/abs/10.1145/3459637.3482104>.
- Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. “Fairness through awareness.” In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226. <https://dl.acm.org/doi/10.1145/2090236.2090255>.
- Elsayed, Gamaleldin, Dilip Krishnan, Hossein Mobahi, Kevin Regan, and Samy Bengio. 2018. “Large margin deep networks for classification.” *Advances in neural information processing systems* 31. https://proceedings.neurips.cc/paper_files/paper/2018/file/42998cf32d552343bc8e460416382dca-Paper.pdf.
- European Parliament and Council of the 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)*. Official Journal L 119, 4.5.2016, p. 1–88. Accessed January 22, 2026. <https://data.europa.eu/eli/reg/2016/679/oj>.
- Ferry, Julien, Ulrich Aivodji, Sébastien Gambs, Marie-José Huguet, and Mohamed Siala. 2023. “Improving fairness generalization through a sample-robust optimization method.” *Machine Learning* 112 (6): 2131–2192. <https://hal.science/hal-03709547>.
- Freiesleben, Timo, and Thomas Grote. 2023. “Beyond generalization: a theory of robustness in machine learning.” *Synthese* 202 (4): 109. <https://doi.org/10.1007/s11229-023-04334-9>.
- Gonen, Hila, and Yoav Goldberg. 2019. “Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them.” In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 609–614. <https://aclanthology.org/N19-1061/>.
- Goodfellow, Ian J, Jonathon Shlens, and Christian Szegedy. 2014. “Explaining and harnessing adversarial examples.” *arXiv preprint arXiv:1412.6572*, <http://arxiv.org/abs/1412.6572>.
- Grari, Vincent, Thibault Laugel, Tatsunori Hashimoto, Marcin Detyniecki, et al. 2024. “On the fairness road: Robust optimization for adversarial debiasing.” In *International Conference on Learning Representations*, 2024:36608–36638. https://proceedings.iclr.cc/paper_files/paper/2024/file/9deeb6f5be09657d9e6e6c0d2416af47-Paper-Conference.pdf.
- Hardt, Moritz, Eric Price, and Nati Srebro. 2016. “Equality of opportunity in supervised learning.” *Advances in neural information processing systems* 29. https://proceedings.neurips.cc/paper_files/paper/2016/file/6a9659feb1216f14f7384ba499518b38-Paper.pdf.
- Islam, Rashidul, Huiyuan Chen, and Yiwei Cai. 2024. “Fairness without demographics through shared latent space-based debiasing.” In *Proceedings of the AAAI Conference on Artificial Intelligence*, 38:12717–12725. 11. <https://ojs.aaai.org/index.php/AAAI/article/view/29167>.
- Jesus, Sérgio, José Pombal, Duarte Alves, André Cruz, Pedro Saleiro, Rita Ribeiro, João Gama, and Pedro Bizarro. 2022. “Turning the tables: Biased, imbalanced, dynamic tabular datasets for ml evaluation.” *Advances in Neural Information Processing Systems* 35:33563–33575. https://proceedings.neurips.cc/paper_files/paper/2022/hash/d9696563856bd350e4e7ac5e5812f23c-Abstract-Datasets_and_Benchmarks.html.

- Jiang, Yiding, Dilip Krishnan, Hossein Mobahi, and Samy Bengio. 2018. "Predicting the generalization gap in deep networks with margin distributions." *arXiv preprint arXiv:1810.00113*, <http://arxiv.org/abs/1810.00113>.
- Jiang, Zhimeng, Xiaotian Han, Hongye Jin, Guanchu Wang, Rui Chen, Na Zou, and Xia Hu. 2023. "Chasing fairness under distribution shift: A model weight perturbation approach." *arXiv preprint arXiv:2303.03300*, <https://arxiv.org/abs/2303.03300>.
- Kaur, Komalpreet, Yogesh Kumar, and Sukhpreet Kaur. 2023. "Artificial intelligence and machine learning in financial services to improve the business system." In *Computational Intelligence for Modern Business Systems: Emerging Applications and Strategies*, 3–30. Springer. https://doi.org/10.1007/978-981-99-5354-7_1.
- Kemp, Steven, and Nieves Erades Pérez. 2023. "Consumer fraud against older adults in digital society: Examining victimization and its impact." *International Journal of Environmental Research and Public Health* 20 (7): 5404. <https://www.mdpi.com/1660-4601/20/7/5404>.
- Kim, Hoki, Jinseong Park, Yujin Choi, and Jaewook Lee. 2023. "Fantastic robustness measures: The secrets of robust generalization." *Advances in Neural Information Processing Systems* 36:48793–48818. https://proceedings.neurips.cc/paper_files/paper/2023/file/98a5c0470e57d518ade4e56c6ee0b363-Paper-Conference.pdf.
- Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. 2016. "Inherent trade-offs in the fair determination of risk scores." *arXiv preprint arXiv:1609.05807*, <http://arxiv.org/abs/1609.05807>.
- Koenecke, Allison, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Touns, John R Rickford, Dan Jurafsky, and Sharad Goel. 2020. "Racial disparities in automated speech recognition." *Proceedings of the national academy of sciences* 117 (14): 7684–7689. <https://www.pnas.org/doi/abs/10.1073/pnas.1915768117>.
- Kuhn, Harold W. 1955. "The Hungarian method for the assignment problem." *Naval research logistics quarterly* 2 (1-2): 83–97. <https://onlinelibrary.wiley.com/doi/abs/10.1002/nav.3800020109>.
- Kusner, Matt J, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. "Counterfactual fairness." *Advances in neural information processing systems* 30. https://proceedings.neurips.cc/paper_files/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.
- Lin, Yujie, Chen Zhao, Minglai Shao, Baoluo Meng, Xujiang Zhao, and Haifeng Chen. 2023. "Towards counterfactual fairness-aware domain generalization in changing environments." *arXiv preprint arXiv:2309.13005*, <https://arxiv.org/pdf/2309.13005>.
- Lipman, Yaron, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2022. "Flow matching for generative modeling." *arXiv preprint arXiv:2210.02747*, <http://arxiv.org/abs/2210.02747>.
- Lyu, Shen-Huan, Lu Wang, and Zhi-Hua Zhou. 2022. "Improving generalization of deep neural networks by leveraging margin distribution." *Neural Networks* 151:48–60. <https://www.sciencedirect.com/science/article/pii/S0893608022000971>.
- Makowski, Maciej. 2024. "Feature Importance Mapping in performative predictions." Master's thesis. <https://studenttheses.uu.nl/items/0a323180-5c03-40cc-8d64-291adebc9256>.

- Mandal, Debmalya, Samuel Deng, Suman Jana, Jeannette Wing, and Daniel J Hsu. 2020. "Ensuring fairness beyond the training data." *Advances in neural information processing systems* 33:18445–18456. https://proceedings.neurips.cc/paper_files/paper/2020/hash/d6539d3b57159babf6a72e106beb45bd-Abstract.html.
- Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. "A survey on bias and fairness in machine learning." *ACM computing surveys (CSUR)* 54 (6): 1–35. <https://dl.acm.org/doi/10.1145/3457607>.
- Mishler, Alan, and Niccolò Dalmaso. 2022. "Fair when trained, unfair when deployed: Observable fairness measures are unstable in performative prediction settings." *arXiv preprint arXiv:2202.05049*, <http://arxiv.org/abs/2202.05049>.
- Moosavi-Dezfooli, Seyed-Mohsen, Alhussein Fawzi, and Pascal Frossard. 2016. "Deepfool: a simple and accurate method to fool deep neural networks." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2574–2582. https://openaccess.thecvf.com/content_cvpr_2016/html/Moosavi-Dezfooli_DeepFool_A_Simple_CVPR_2016_paper.html.
- Nanda, Vedant, Samuel Dooley, Sahil Singla, Soheil Feizi, and John P Dickerson. 2021. "Fairness through robustness: Investigating robustness disparity in deep learning." In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 466–477. <https://dl.acm.org/doi/10.1145/3442188.3445910>.
- O'Connor, Sinead, and Helen Liu. 2024. "Gender bias perpetuation and mitigation in AI technologies: challenges and opportunities." *AI & SOCIETY* 39 (4): 2045–2057. <https://doi.org/10.1007/s00146-023-01675-4>.
- Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. "Dissecting racial bias in an algorithm used to manage the health of populations." *Science* 366 (6464): 447–453. <https://www.science.org/doi/abs/10.1126/science.aax2342>.
- Paudice, Andrea, Luis Muñoz-González, and Emil C Lupu. 2018. "Label sanitization against label flipping poisoning attacks." In *Joint European conference on machine learning and knowledge discovery in databases*, 5–15. Springer. https://doi.org/10.1007/978-3-030-13453-2_1.
- Perdomo, Juan, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. 2020. "Performative prediction." In *International Conference on Machine Learning*, 7599–7609. PMLR. <https://proceedings.mlr.press/v119/perdomo20a.html>.
- Rajan, Anurag, Joel Janai, Andreas Geiger, and Michael J Black. 2019. "Attacking optical flow." In *Proceedings of the IEEE/CVF international conference on computer vision*, 2404–2413. https://openaccess.thecvf.com/content_ICCV_2019/html/Ranjan_Attacking_Optical_Flow_ICCV_2019_paper.html.
- Rajkomar, Alvin, Jeffrey Dean, and Isaac Kohane. 2019. "Machine learning in medicine." *New England Journal of Medicine* 380 (14): 1347–1358. <https://www.nejm.org/doi/full/10.1056/NEJMr1814259>.
- Rezaei, Ashkan, Rizal Fathony, Omid Memarrast, and Brian Ziebart. 2020. "Fairness for robust log loss classification." In *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:5511–5518. 04. <https://ojs.aaai.org/index.php/AAAI/article/view/6002>.

- Roh, Yuji, Kangwook Lee, Steven Whang, and Changho Suh. 2020. “Fr-train: A mutual information-based approach to fair and robust training.” In *International Conference on Machine Learning*, 8147–8157. PMLR. <https://proceedings.mlr.press/v119/roh20a.html>.
- Saleiro, Pedro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T Rodolfa, and Rayid Ghani. 2018. “Aequitas: A bias and fairness audit toolkit.” *arXiv preprint arXiv:1811.05577*, <http://arxiv.org/abs/1811.05577>.
- Salman, Hadi, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sebastien Bubeck, and Greg Yang. 2019. “Provably robust deep learning via adversarially trained smoothed classifiers.” *Advances in neural information processing systems* 32. https://proceedings.neurips.cc/paper_files/paper/2019/file/3a24b25a7b092a252166a1641ae953e7-Paper.pdf.
- Sehwag, Vikash, Arjun Nitin Bhagoji, Liwei Song, Chawin Sitawarin, Daniel Cullina, Mung Chiang, and Prateek Mittal. 2019. “Analyzing the robustness of open-world machine learning.” In *Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security*, 105–116. <https://dl.acm.org/doi/abs/10.1145/3338501.3357372>.
- Shalev-Shwartz, Shai. 2007. *Online learning: Theory, algorithms, and applications*. Hebrew University. <https://home.ttic.edu/~shai/papers/ShalevThesis07.pdf>.
- Sharma, Shubham, Alan H Gee, David Paydarfar, and Joydeep Ghosh. 2021. “Fair-n: Fair and robust neural networks for structured data.” In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 946–955. <https://dl.acm.org/doi/10.1145/3461702.3462559>.
- Stevenson, Megan. 2018. “Assessing risk assessment in action.” *Minn. L. Rev.* 103:303. <https://papers.ssrn.com/abstract=3016088>.
- Suresh, Harini, and John V Guttag. 2019. “A framework for understanding unintended consequences of machine learning.” *arXiv preprint arXiv:1901.10002* 2 (8): 73. <http://arxiv.org/abs/1901.10002>.
- Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. “Intriguing properties of neural networks.” *arXiv preprint arXiv:1312.6199*, <http://arxiv.org/abs/1312.6199>.
- Taori, Rohan, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. 2020. “Measuring robustness to natural distribution shifts in image classification.” *Advances in Neural Information Processing Systems* 33:18583–18599. https://proceedings.neurips.cc/paper_files/paper/2020/file/d8330f857a17c53d217014ee776bfd50-Paper.pdf.
- Tong, Alexander, Kilian ... Fatras, and Yoshua Bengio. 2023. “Improving and generalizing flow-based generative models with minibatch optimal transport.” *arXiv preprint arXiv:2302.00482*, <http://arxiv.org/abs/2302.00482>.
- Tong, Alexander, Nikolay Malkin, Guillaume Hugué, Yanlei Zhang, Jarrid Rector-Brooks, Kilian Fatras, Guy Wolf, and Yoshua Bengio. 2023. “Conditional flow matching: Simulation-free dynamic optimal transport.” *arXiv preprint arXiv:2302.00482* 2 (3). <https://arxiv.org/abs/2302.00482>.
- Wilson, Christo, Avijit Ghosh, Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary, Kelly Trindel, and Frida Polli. 2021. “Building and auditing fair algorithms: A case study in candidate screening.” In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 666–677. <https://dl.acm.org/doi/10.1145/3442188.3445928>.

- Ying, Wangyang, Nanxu Gong, Dongjie Wang, Xinyuan Wang, Arun Vignesh Malarkkan, Vivek Gupta, Chandan K Reddy, and Yanjie Fu. 2025. "Distribution shift aware neural tabular learning." *arXiv preprint arXiv:2508.19486*, <https://arxiv.org/pdf/2508.19486>.
- Yurtsever, Ekim, Jacob Lambert, Alexander Carballo, and Kazuya Takeda. 2020. "A survey of autonomous driving: Common practices and emerging technologies." *IEEE access* 8:58443–58469. <https://ieeexplore.ieee.org/abstract/document/9046805>.
- Završnik, Aleš. 2021. "Algorithmic justice: Algorithms and big data in criminal justice settings." *European Journal of criminology* 18 (5): 623–642. <https://doi.org/10.1177/1477370819876762>.
- Zech, John R, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. 2018. "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study." *PLoS medicine* 15 (11): e1002683. <https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002683>.
- Zhang, Chiyuan, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2016. "Understanding deep learning requires rethinking generalization." *arXiv preprint arXiv:1611.03530*, <http://arxiv.org/abs/1611.03530>.

Appendix A

Supplemental Material

A.1 Multi-Objective Ranking Formalization

To compute the *Weighted Rank* presented in Chapter 6, we utilize a multi-objective scoring function that balances predictive utility, algorithmic fairness, and geometric boundary robustness. The final weighted rank R_w for a given method is defined as the weighted average of its individual metric ranks r_i :

$$R_w = \sum_{i \in \mathcal{M}} w_i \cdot r_i, \quad (\text{A.1})$$

where $\mathcal{M}$ represents the set of evaluation metrics and w_i are the normalized weights. The weighting schema applied to the consolidated ranking matrix is designed to ensure that each conceptual evaluation dimension contributes equally to the final Weighted Rank, regardless of the number of individual metrics representing it. Specifically, the metrics are partitioned into three groups: predictive utility (*Accuracy* and *F1-score*), traditional group fairness (**DP**, **TP** parity, and **FP** parity), and geometric boundary robustness (*Recourse Burden*, and the *TabMIST* composite). Within each group, individual metric ranks are averaged, and the three resulting group scores are then averaged equally to produce the final Weighted Rank. This grouped averaging prevents any single conceptual dimension from dominating the ranking simply by virtue of having more constituent metrics, a distortion that would arise under naive equal-weight averaging, where the three fairness metrics would collectively account for 37.5% of the total score. *TabMISTAvg* and *TabMISTStd* are treated as a single composite point within the geometric group, following the original *tsMIST* formulation in which these two statistics are understood as complementary descriptors of the same underlying boundary distribution rather than independent measurements. The resulting weight distribution, summarized in Table A.1, reflects this three-way partition precisely.

Table A.1: Weighting schema applied to the consolidated metric ranking matrix.

Metric	Weight (w_i)
Accuracy	0.1667
F_1 -Score	0.1667
Demographic Parity (DP)	0.1111
Equalized Odds (TP)	0.1111
Equalized Odds (FP)	0.1111
Recourse Burden	0.1667
$TabMIST_{Avg}$ (M_{Mean})	0.0833
$TabMIST_{Std}$ (M_{Std})	0.0833

A.2 Directional Categories and Evaluation Framework

This section provides documentation of the directional metric decay trajectories (Δ) observed over the course of the dynamic optimization horizon across all tested datasets. The temporal delta is formally computed as $\Delta = \text{Final} - \text{Initial}$. This detailed breakdown serves as the quantitative foundation for the high-level synthesis and research question evaluations presented in Chapter 6.

Evaluating these temporal trajectories requires distinguishing between two distinct directional optimization categories:

1. **Utility and Group Equality Ratios (Higher is Better):** For Accuracy, F_1 -score, Demographic Parity (**DP**) Ratio, and **TP** Ratio, a positive delta ($\Delta > 0$) indicates compounding optimization or stabilization over time, while a negative delta ($\Delta < 0$) signifies structural decay.
2. **Error Disparities and Systemic Cost (Lower is Better):** For **FP** Ratio and Recourse Burden, a negative delta ($\Delta < 0$) represents successful minimization of demographic penalties, whereas a positive delta ($\Delta > 0$) points to a compounding accumulation of bias or operational cost.

Crucially, the *TabMIST* metrics (M_{Mean} and M_{Std}) measure absolute coordinate positioning along a continuous interpolation trajectory. The ideal state for M_{Mean} is 0.5, representing a balanced decision boundary centered precisely between demographic cohorts. A discrepancy from 0.5 in the raw M_{Mean} values, or a non-zero ΔM_{Mean} , indicates a “boundary drift” where the classifier’s decision surface has skewed toward one specific demographic group, signifying a loss of spatial neutrality. Results are presented in Tables B.13 through B.18.

Appendix B

Tables of Results

Table B.1: Folktable Dataset Results: Comparison between different bias mitigation approaches and FairMorph augmentation. Results averaged across training steps, where original ground truth was used (BASELINE). Values in bold highlight the best-performing method regarding accuracy, F1-score, and the ratio of fairness metrics across groups.

Method	Group	Acc $\uparrow$	F1 $\uparrow$	DP $\uparrow$	TPR $\uparrow$	FPR $\downarrow$	Burden $\downarrow$	M_Mean _{50%}	M_Std $\downarrow$
FairMorph _{1epoch}	0	0.778	0.778	0.547	0.798	0.258	2.134	53.573	26.699
	1	0.778	0.778	0.410	0.756	0.183	2.462	58.232	25.592
	Ratio	-	-	0.753	0.948	0.733	0.867	0.921	0.959
FairMorph _{5epoch}	0	0.781	0.783	0.560	0.814	0.268	2.061	53.484	26.220
	1	0.781	0.783	0.381	0.724	0.157	2.336	59.485	25.149
	Ratio	-	-	0.677	0.886	0.578	0.882	0.899	0.959
FaiR-N _(0.5,0.5)	0	0.732	0.673	0.399	0.628	0.136	1.092	59.741	26.547
	1	0.732	0.673	0.136	0.313	0.023	1.096	74.737	21.448
	Ratio	-	-	0.331	0.483	0.160	0.982	0.800	0.808
FaiR-N _(1,0)	0	0.660	0.511	0.277	0.456	0.071	0.430	62.550	25.570
	1	0.660	0.511	0.005	0.012	0.001	0.425	51.127	5.011
	Ratio	-	-	0.017	0.024	0.013	0.990	0.836	0.595
Mixup DP	0	0.758	0.737	0.444	0.678	0.173	1.788	56.585	26.641
	1	0.758	0.737	0.395	0.733	0.173	1.889	55.540	25.658
	Ratio	-	-	0.887	0.920	0.850	0.945	0.966	0.953
Mixup EO	0	0.762	0.754	0.503	0.740	0.230	1.374	54.276	26.402
	1	0.762	0.754	0.422	0.763	0.200	1.426	54.115	25.433
	Ratio	-	-	0.844	0.967	0.853	0.954	0.971	0.960
Vanilla	0	0.777	0.773	0.536	0.787	0.247	2.103	54.572	26.290
	1	0.777	0.773	0.341	0.674	0.124	2.437	61.689	24.624
	Ratio	-	-	0.634	0.853	0.496	0.864	0.886	0.936

Table B.2: Folktable Dataset Results: Comparison between different bias mitigation approaches and FairMorph augmentation. Results averaged across training steps, where model predictions were used as labels (PERFORMATIVE). Values in bold highlight the best-performing method regarding accuracy, F1-score, and the ratio of fairness metrics across groups.

Method	Group	Acc $\uparrow$	F1 $\uparrow$	DP $\uparrow$	TPR $\uparrow$	FPR $\downarrow$	Burden $\downarrow$	M_Mean _{50%}	M_Std $\downarrow$
FairMorph _{1epoch}	0	0.766	0.754	0.503	0.746	0.224	3.326	54.757	26.319
	1	0.766	0.754	0.380	0.711	0.164	3.724	60.078	25.023
	Ratio	-	-	0.770	0.934	0.742	0.888	0.912	0.951
FairMorph _{5epoch}	0	0.777	0.778	0.559	0.810	0.270	3.217	53.188	26.611
	1	0.777	0.778	0.396	0.734	0.175	3.600	58.069	25.396
	Ratio	-	-	0.709	0.905	0.661	0.885	0.917	0.953
FaiR-N _(0.5,0.5)	0	0.681	0.555	0.314	0.504	0.096	1.659	63.277	25.501
	1	0.681	0.555	0.066	0.148	0.013	1.649	62.888	15.582
	Ratio	-	-	0.171	0.246	0.097	0.987	0.825	0.795
FaiR-N _(1,0)	0	0.612	0.378	0.185	0.310	0.040	1.260	67.229	23.046
	1	0.612	0.378	0.005	0.010	0.001	1.248	16.127	4.211
	Ratio	-	-	0.014	0.020	0.009	0.986	0.950	0.963
Mixup DP	0	0.749	0.720	0.434	0.661	0.172	2.916	56.858	26.266
	1	0.749	0.720	0.403	0.733	0.188	3.071	55.136	25.085
	Ratio	-	-	0.898	0.890	0.817	0.942	0.954	0.952
Mixup EO	0	0.742	0.711	0.454	0.673	0.201	2.224	55.511	26.174
	1	0.742	0.711	0.394	0.712	0.187	2.307	55.819	24.372
	Ratio	-	-	0.870	0.932	0.823	0.953	0.969	0.932
Vanilla	0	0.764	0.746	0.498	0.741	0.218	2.969	56.038	26.321
	1	0.764	0.746	0.290	0.586	0.097	3.509	65.131	23.957
	Ratio	-	-	0.574	0.779	0.435	0.843	0.860	0.910

Table B.3: Account Fraud Dataset Results: Comparison between different bias mitigation approaches and FairMorph augmentation. Results averaged across training steps, where original ground truth was used (BASELINE). Values in bold highlight the best-performing method regarding accuracy, F1-score, and the ratio of fairness metrics across groups.

Method	Group	Acc $\uparrow$	F1 $\uparrow$	DP $\uparrow$	TPR $\uparrow$	FPR $\downarrow$	Burden $\downarrow$	M_Mean _{50%}	M_Std $\downarrow$
FairMorph _{1epoch}	0	0.827	0.669	0.252	0.645	0.107	1.276	61.005	24.514
	1	0.827	0.669	0.578	0.778	0.264	1.650	57.590	23.979
	Ratio	-	-	0.443	0.834	0.489	0.789	0.896	0.848
FairMorph _{5epoch}	0	0.822	0.658	0.249	0.631	0.108	1.210	60.836	24.308
	1	0.822	0.658	0.646	0.821	0.369	1.808	52.252	25.282
	Ratio	-	-	0.391	0.770	0.328	0.679	0.829	0.911
FaiR-N _(0.5,0.5)	0	0.750	0.189	0.035	0.104	0.010	0.005	34.223	7.127
	1	0.750	0.189	0.140	0.205	0.035	0.001	0.000	0.000
	Ratio	-	-	0.279	0.390	0.125	0.137	0.429	0.429
FaiR-N _(1,0)	0	0.783	0.386	0.116	0.308	0.043	0.004	26.498	10.448
	1	0.783	0.386	0.339	0.435	0.193	0.001	19.299	10.528
	Ratio	-	-	0.310	0.691	0.201	0.425	0.312	0.382
Mixup DP	0	0.788	0.578	0.270	0.598	0.146	0.550	56.545	25.772
	1	0.788	0.578	0.397	0.532	0.183	0.566	55.648	21.238
	Ratio	-	-	0.658	0.891	0.600	0.806	0.871	0.735
Mixup EO	0	0.663	0.535	0.439	0.685	0.343	0.200	51.858	26.024
	1	0.663	0.535	0.595	0.712	0.398	0.210	42.025	25.541
	Ratio	-	-	0.742	0.866	0.742	0.915	0.809	0.824
Vanilla	0	0.825	0.655	0.235	0.611	0.097	1.224	61.793	24.279
	1	0.825	0.655	0.654	0.841	0.357	2.017	49.759	22.909
	Ratio	-	-	0.363	0.726	0.297	0.618	0.805	0.923

Table B.4: Account Fraud Dataset Results: Comparison between different bias mitigation approaches and FairMorph augmentation. Results averaged across training steps, where model predictions were used as labels (PERFORMATIVE). Values in bold highlight the best-performing method regarding accuracy, F1-score, and the ratio of fairness metrics across groups.

Method	Group	Acc $\uparrow$	F1 $\uparrow$	DP $\uparrow$	TPR $\uparrow$	FPR $\downarrow$	Burden $\downarrow$	M_Mean _{50%}	M_Std $\downarrow$
FairMorph _{1epoch}	0	0.810	0.661	0.289	0.684	0.144	5.235	56.565	25.190
	1	0.810	0.661	0.602	0.762	0.348	5.448	48.943	24.839
	Ratio	-	-	0.489	0.905	0.452	0.826	0.860	0.818
FairMorph _{5epoch}	0	0.804	0.659	0.301	0.695	0.156	5.241	55.106	25.160
	1	0.804	0.659	0.655	0.818	0.392	6.587	49.868	21.572
	Ratio	-	-	0.465	0.852	0.419	0.759	0.804	0.815
FaiR-N _(0.5,0.5)	0	0.674	0.188	0.172	0.213	0.158	0.119	45.897	4.364
	1	0.674	0.188	0.210	0.242	0.152	0.178	12.857	0.000
	Ratio	-	-	0.466	0.584	0.167	0.144	0.528	0.429
FaiR-N _(1,0)	0	0.754	0.214	0.038	0.119	0.009	0.005	0.000	0.000
	1	0.754	0.214	0.159	0.235	0.036	0.019	0.000	0.000
	Ratio	-	-	0.251	0.517	0.106	0.020	0.000	0.000
Mixup DP	0	0.736	0.159	0.041	0.092	0.022	0.041	49.610	3.721
	1	0.736	0.159	0.058	0.080	0.020	0.016	8.571	0.000
	Ratio	-	-	0.512	0.672	0.266	0.076	0.556	0.000
Mixup EO	0	0.546	0.320	0.403	0.414	0.399	0.316	52.655	27.142
	1	0.546	0.320	0.469	0.464	0.482	0.352	57.095	23.602
	Ratio	-	-	0.856	0.904	0.818	0.866	0.794	0.774
Vanilla	0	0.812	0.662	0.285	0.677	0.140	4.839	57.006	24.861
	1	0.812	0.662	0.662	0.830	0.392	6.958	47.633	21.549
	Ratio	-	-	0.431	0.816	0.377	0.673	0.785	0.790

Table B.5: IBM Transaction Fraud Dataset Results: Comparison between different bias mitigation approaches and FairMorph augmentation. Results averaged across training steps, where original ground truth was used (BASELINE). Values in bold highlight the best-performing method regarding accuracy, F1-score, and the ratio of fairness metrics across groups.

Method	Group	Acc $\uparrow$	F1 $\uparrow$	DP $\uparrow$	TPR $\uparrow$	FPR $\downarrow$	Burden $\downarrow$	M_Mean _{50%}	M_Std $\downarrow$
FairMorph _{1epoch}	0	0.755	0.540	0.298	0.493	0.141	1.821	54.467	18.255
	1	0.755	0.540	0.332	0.497	0.141	1.941	53.533	18.372
	Ratio	-	-	0.794	0.905	0.772	0.745	0.826	0.808
FairMorph _{5epoch}	0	0.744	0.540	0.316	0.500	0.173	1.830	55.067	18.262
	1	0.744	0.540	0.362	0.524	0.176	1.947	54.439	17.542
	Ratio	-	-	0.746	0.833	0.696	0.748	0.802	0.802
FaiR-N _(0.5,0.5)	0	0.539	0.627	0.808	0.951	0.683	0.766	31.883	12.909
	1	0.539	0.627	0.873	0.975	0.741	0.816	24.391	12.545
	Ratio	-	-	0.912	0.975	0.877	0.958	0.443	0.548
FaiR-N _(1,0)	0	0.681	0.495	0.468	0.593	0.374	0.100	30.353	9.879
	1	0.681	0.495	0.522	0.604	0.436	0.098	41.890	9.051
	Ratio	-	-	0.736	0.848	0.674	0.683	0.599	0.555
Mixup DP	0	0.734	0.520	0.310	0.476	0.181	1.415	50.024	16.936
	1	0.734	0.520	0.367	0.526	0.187	1.608	47.545	17.309
	Ratio	-	-	0.760	0.849	0.783	0.747	0.805	0.813
Mixup EO	0	0.596	0.634	0.688	0.876	0.548	1.527	40.119	22.073
	1	0.596	0.634	0.756	0.916	0.578	1.627	33.442	18.958
	Ratio	-	-	0.896	0.952	0.933	0.938	0.952	0.969
Vanilla	0	0.759	0.551	0.310	0.511	0.152	1.799	55.584	17.694
	1	0.759	0.551	0.361	0.534	0.162	1.875	54.679	17.847
	Ratio	-	-	0.732	0.835	0.678	0.755	0.803	0.779

Table B.6: IBM Transaction Fraud Dataset Results: Comparison between different bias mitigation approaches and FairMorph augmentation. Results averaged across training steps, where model predictions were used as labels (PERFORMATIVE). Values in bold highlight the best-performing method regarding accuracy, F1-score, and the ratio of fairness metrics across groups.

Method	Group	Acc $\uparrow$	F1 $\uparrow$	DP $\uparrow$	TPR $\uparrow$	FPR $\downarrow$	Burden $\downarrow$	M_Mean _{50%}	M_Std $\downarrow$
FairMorph _{1epoch}	0	0.573	0.048	0.037	0.027	0.046	1.470	41.886	7.490
	1	0.573	0.048	0.039	0.030	0.054	1.753	24.766	10.072
	Ratio	-	-	0.713	0.640	0.664	0.514	0.542	0.360
FairMorph _{5epoch}	0	0.572	0.041	0.034	0.025	0.044	1.612	41.679	15.489
	1	0.572	0.041	0.032	0.024	0.047	1.531	33.465	12.108
	Ratio	-	-	0.687	0.591	0.666	0.596	0.575	0.417
FaiR-N _(0.5,0.5)	0	0.409	0.563	1.000	1.000	1.000	3.360	0.000	0.000
	1	0.409	0.563	1.000	1.000	1.000	3.373	0.000	0.000
	Ratio	-	-	1.000	1.000	1.000	0.995	0.000	0.000
FaiR-N _(1,0)	0	0.577	0.066	0.043	0.041	0.050	0.002	11.571	2.903
	1	0.577	0.066	0.035	0.035	0.043	0.004	12.476	1.136
	Ratio	-	-	0.721	0.747	0.606	0.204	0.132	0.056
Mixup DP	0	0.580	0.068	0.034	0.040	0.038	0.745	31.813	4.785
	1	0.580	0.068	0.036	0.041	0.044	0.523	36.837	6.582
	Ratio	-	-	0.783	0.757	0.741	0.417	0.393	0.116
Mixup EO	0	0.412	0.564	0.996	0.998	0.996	18.336	5.810	3.033
	1	0.412	0.564	0.997	0.998	0.997	18.749	0.000	0.000
	Ratio	-	-	0.999	1.000	0.999	0.976	0.143	0.143
Vanilla	0	0.574	0.049	0.043	0.036	0.053	1.698	29.562	14.359
	1	0.574	0.049	0.020	0.019	0.029	1.223	37.114	13.456
	Ratio	-	-	0.498	0.507	0.513	0.475	0.455	0.511

Table B.7: Folktable Dataset Metric Rankings. Ranked over the average of the fairness ratios and performance metrics for the different methods. (BASELINE)

Method	Acc	F1	DP	TPR	FPR	Burden	M_Mean	M_Std	Weighted Rank
FairMorph _{1epoch}	2.0	2.0	3.0	2.0	3.0	6.0	3.0	3.0	3.06
FairMorph _{5epoch}	1.0	1.0	4.0	4.0	4.0	5.0	4.0	2.0	3.0
FaiR-N _(0.5,0.5)	6.0	6.0	6.0	6.0	6.0	2.0	7.0	6.0	5.42
FaiR-N _(1,0)	7.0	7.0	7.0	7.0	7.0	1.0	6.0	7.0	5.92
Mixup DP	5.0	5.0	1.0	3.0	2.0	4.0	2.0	4.0	3.5
Mixup EO	4.0	4.0	2.0	1.0	1.0	3.0	1.0	1.0	2.44
Vanilla	3.0	3.0	5.0	5.0	5.0	7.0	5.0	5.0	4.67

Table B.8: Folktable Dataset Metric Rankings. Ranked over the average of the fairness ratios and performance metrics for the different methods. (PERFORMATIVE)

Method	Acc	F1	DP	TPR	FPR	Burden	M_Mean	M_Std	Weighted Rank
FairMorph _{1epoch}	2.0	2.0	3.0	1.0	3.0	5.0	5.0	4.0	3.03
FairMorph _{5epoch}	1.0	1.0	4.0	3.0	4.0	6.0	4.0	2.0	3.06
FaiR-N _(0.5,0.5)	6.0	6.0	6.0	6.0	6.0	1.0	7.0	7.0	5.33
FaiR-N _(1,0)	7.0	7.0	7.0	7.0	7.0	2.0	3.0	1.0	5.33
Mixup DP	4.0	4.0	1.0	4.0	2.0	4.0	2.0	3.0	3.19
Mixup EO	5.0	5.0	2.0	2.0	1.0	3.0	1.0	5.0	3.22
Vanilla	3.0	3.0	5.0	5.0	5.0	7.0	6.0	6.0	4.83

Table B.9: Account Fraud Dataset Metric Rankings. Ranked over the average of the fairness ratios and performance metrics for the different methods. (BASELINE)

Method	Acc	F1	DP	TPR	FPR	Burden	M_Mean	M_Std	Weighted Rank
FairMorph _{1epoch}	1.0	1.0	3.0	3.0	3.0	3.0	1.0	3.0	2.17
FairMorph _{5epoch}	3.0	2.0	4.0	4.0	4.0	4.0	3.0	2.0	3.25
FaiR-N _(0.5,0.5)	6.0	7.0	7.0	7.0	7.0	7.0	6.0	6.0	6.67
FaiR-N _(1,0)	5.0	6.0	6.0	6.0	6.0	6.0	7.0	7.0	6.0
Mixup DP	4.0	4.0	2.0	1.0	2.0	2.0	2.0	5.0	2.81
Mixup EO	7.0	5.0	1.0	2.0	1.0	1.0	4.0	4.0	3.28
Vanilla	2.0	3.0	5.0	5.0	5.0	5.0	5.0	1.0	3.83

Table B.10: Account Fraud Dataset Metric Rankings. Ranked over the average of the fairness ratios and performance metrics for the different methods. (PERFORMATIVE)

Method	Acc	F1	DP	TPR	FPR	Burden	M_Mean	M_Std	Weighted Rank
FairMorph _{1epoch}	2.0	2.0	3.0	1.0	2.0	2.0	1.0	1.0	1.83
FairMorph _{5epoch}	3.0	3.0	5.0	3.0	3.0	3.0	2.0	2.0	3.06
FaiR-N _(0.5,0.5)	6.0	6.0	4.0	6.0	6.0	5.0	6.0	5.0	5.53
FaiR-N _(1,0)	4.0	5.0	7.0	7.0	7.0	7.0	7.0	6.0	6.08
Mixup DP	5.0	7.0	2.0	5.0	5.0	6.0	5.0	6.0	5.25
Mixup EO	7.0	4.0	1.0	2.0	1.0	1.0	3.0	4.0	3.03
Vanilla	1.0	1.0	6.0	4.0	4.0	4.0	4.0	3.0	3.14

Table B.11: IBM Transaction Fraud Dataset Metric Rankings. Ranked over the average of the fairness ratios and performance metrics for the different methods. (BASELINE)

Method	Acc	F1	DP	TPR	FPR	Burden	M_Mean	M_Std	Weighted Rank
FairMorph _{1epoch}	2.0	4.0	3.0	3.0	4.0	6.0	2.0	3.0	3.53
FairMorph _{5epoch}	3.0	5.0	5.0	7.0	5.0	4.0	5.0	4.0	4.64
FaiR-N _(0.5,0.5)	7.0	2.0	1.0	1.0	2.0	1.0	7.0	7.0	3.28
FaiR-N _(1,0)	5.0	7.0	6.0	5.0	7.0	7.0	6.0	6.0	6.17
Mixup DP	4.0	6.0	4.0	4.0	3.0	5.0	3.0	2.0	4.14
Mixup EO	6.0	1.0	2.0	2.0	1.0	2.0	1.0	1.0	2.22
Vanilla	1.0	3.0	7.0	6.0	6.0	3.0	4.0	5.0	4.03

Table B.12: IBM Transaction Fraud Dataset Metric Rankings. Ranked over the average of the fairness ratios and performance metrics for the different methods. FaiR-N_{0.5,0.5} and Mixup DP methods are not included since the model only predicts one class, therefore being useless. (PERFORMATIVE)

Method	Acc	F1	DP	TPR	FPR	Burden	M_Mean	M_Std	Weighted Rank
FairMorph _{1epoch}	4.0	4.0	3.0	3.0	3.0	2.0	2.0	3.0	3.08
FairMorph _{5epoch}	5.0	5.0	4.0	4.0	2.0	1.0	1.0	2.0	3.19
FaiR-N _(0.5,0.5)	-	-	-	-	-	-	-	-	-
FaiR-N _(1,0)	2.0	2.0	2.0	2.0	4.0	5.0	5.0	5.0	3.22
Mixup DP	1.0	1.0	1.0	1.0	1.0	4.0	4.0	4.0	2.0
Mixup EO	-	-	-	-	-	-	-	-	-
Vanilla	3.0	3.0	5.0	5.0	5.0	3.0	3.0	1.0	3.5

Table B.13: Folktable Dataset metric decay over time: Difference in results from the last training step and the first training step, where original ground truth was used (BASELINE). Values in bold highlight the lowest decay across methods regarding accuracy, F1-score, and the ratio of fairness metrics across groups.

Method	Acc $\uparrow$	F1 $\uparrow$	DP $\uparrow$	TPR $\uparrow$	FPR $\downarrow$	Burden $\downarrow$	M_Mean (50%)	M_Std $\downarrow$
FairMorph _{1epoch}	-0.026	-0.065	0.145	0.078	0.361	0.021	0.123	0.033
FairMorph _{5epoch}	-0.018	-0.050	-0.054	-0.103	-0.068	0.079	0.014	0.023
FaiR-N _(0.5,0.5)	-0.071	-0.160	-0.260	-0.383	-0.192	0.008	-0.025	-0.120
FaiR-N _(1,0)	-0.098	-0.272	-0.072	-0.102	-0.045	0.008	0.252	0.184
Mixup DP	-0.046	-0.116	0.131	-0.168	-0.065	0.126	0.006	0.069
Mixup EO	-0.039	-0.098	0.066	-0.027	0.171	0.107	0.008	0.070
Vanilla	-0.019	-0.049	0.018	-0.033	0.038	0.101	0.075	0.038

Table B.14: Folktable Dataset metric decay over time: Difference in results from the last training step and the first training step, where model predictions were used as labels (PERFORMATIVE). Values in bold highlight the lowest decay across methods regarding accuracy, F1-score, and the ratio of fairness metrics across groups.

Method	Acc $\uparrow$	F1 $\uparrow$	DP $\uparrow$	TPR $\uparrow$	FPR $\downarrow$	Burden $\downarrow$	M_Mean (50%)	M_Std $\downarrow$
FairMorph _{1epoch}	-0.070	-0.147	0.237	0.032	0.329	0.124	0.111	0.036
FairMorph _{5epoch}	-0.028	-0.070	0.063	-0.008	0.212	0.111	0.087	0.027
FaiR-N _(0.5,0.5)	-0.177	-0.461	-0.411	-0.589	-0.251	0.013	0.191	0.182
FaiR-N _(1,0)	-0.152	-0.457	-0.072	-0.102	-0.045	-0.016	0.252	0.184
Mixup DP	-0.086	-0.205	0.035	-0.243	-0.219	0.113	-0.019	0.069
Mixup EO	-0.107	-0.258	0.117	-0.145	-0.158	0.125	0.011	0.072
Vanilla	-0.065	-0.144	-0.131	-0.221	-0.100	0.056	0.061	-0.028

Table B.15: Account Fraud Dataset metric decay over time: Difference in results from the last training step and the first training step, where original ground truth was used (BASELINE). Values in bold highlight the lowest decay across methods regarding accuracy, F1-score, and the ratio of fairness metrics across groups.

Method	Acc $\uparrow$	F1 $\uparrow$	DP $\uparrow$	TPR $\uparrow$	FPR $\downarrow$	Burden $\downarrow$	M_Mean (50%)	M_Std $\downarrow$
FairMorph _{1epoch}	0.014	0.077	0.048	0.162	-0.107	0.240	-0.073	0.035
FairMorph _{5epoch}	0.030	0.086	-0.049	0.055	-0.273	0.248	-0.049	0.059
FaiR-N _(0.5,0.5)	-0.035	0.057	0.849	0.589	0.000	0.286	0.000	0.000
FaiR-N _(1,0)	-0.040	-0.008	0.020	-0.031	0.056	0.000	0.000	0.000
Mixup DP	0.044	0.444	0.164	-0.098	0.383	0.276	-0.015	0.884
Mixup EO	0.258	0.293	-0.191	0.125	-0.060	0.073	0.155	-0.404
Vanilla	0.015	0.084	0.007	0.090	-0.163	0.115	0.041	-0.084

Table B.16: Account Fraud Dataset metric decay over time: Difference in results from the last training step and the first training step, where model predictions were used as labels (PERFORMATIVE). Values in bold highlight the lowest decay across methods regarding accuracy, F1-score, and the ratio of fairness metrics across groups.

Method	Acc $\uparrow$	F1 $\uparrow$	DP $\uparrow$	TPR $\uparrow$	FPR $\downarrow$	Burden $\downarrow$	M_Mean (50%)	M_Std $\downarrow$
FairMorph _{1epoch}	0.003	0.065	0.091	0.254	-0.139	0.098	-0.028	-0.204
FairMorph _{5epoch}	0.019	0.078	-0.020	0.074	-0.221	0.380	-0.235	-0.308
FaiR-N _(0.5,0.5)	-0.074	-0.057	0.596	0.444	0.000	0.000	0.000	-1.000
FaiR-N _(1,0)	-0.033	0.010	-0.018	-0.068	-0.019	0.140	0.000	0.000
Mixup DP	-0.052	-0.062	-0.008	-0.300	-0.264	-0.530	-0.108	0.000
Mixup EO	0.119	-0.110	-0.120	0.186	-0.467	0.061	-0.079	-0.998
Vanilla	-0.000	0.068	0.034	0.097	-0.098	0.189	0.003	-0.505

Table B.17: IBM Transaction Fraud Dataset metric decay over time: Difference in results from the last training step and the first training step, where original ground truth was used (BASELINE). Values in bold highlight the lowest decay across methods regarding accuracy, F1-score, and the ratio of fairness metrics across groups.

Method	Acc	F1	DP	TPR	FPR	Burden	M_Mean	M_Std
FairMorph _{1epoch}	-0.116	0.539	0.332	0.099	0.572	0.977	0.001	0.977
FairMorph _{5epoch}	-0.058	0.620	0.315	0.198	0.445	0.923	0.014	0.992
FaiR-N _(0.5,0.5)	0.691	0.607	-0.268	-0.099	-0.409	-0.042	0.282	0.973
FaiR-N _(1,0)	-0.071	0.665	0.438	0.423	0.351	0.980	0.276	0.979
Mixup DP	-0.210	0.313	-0.125	-0.204	0.106	0.821	0.029	0.993
Mixup EO	0.628	0.530	-0.296	-0.215	-0.114	-0.089	-0.102	-0.065
Vanilla	-0.014	0.669	0.414	0.340	0.531	0.895	0.038	0.941

Table B.18: IBM Transaction Fraud Dataset metric decay over time: Difference in results from the last training step and the first training step, where model predictions were used as labels (PERFORMATIVE). Values in bold highlight the lowest decay across methods regarding accuracy, F1-score, and the ratio of fairness metrics across groups.

Method	Acc	F1	DP	TPR	FPR	Burden	M_Mean	M_Std
FairMorph _{1epoch}	-0.368	-0.053	0.547	-0.070	0.630	0.965	0.192	0.997
FairMorph _{5epoch}	-0.371	-0.113	0.446	-0.297	0.452	0.989	0.417	0.724
FaiR-N _(0.5,0.5)	0.367	0.432	0.000	0.000	0.000	-0.004	0.000	0.000
FaiR-N _(1,0)	-0.374	-0.119	0.471	-0.029	0.551	0.000	0.000	0.000
Mixup DP	-0.377	-0.180	0.076	-0.055	0.057	0.573	0.000	0.000
Mixup EO	0.349	0.431	0.007	0.003	0.009	0.020	0.000	-1.000
Vanilla	-0.381	-0.126	0.376	-0.235	0.703	0.873	0.177	0.978