Intersectional Fairness in Healthcare AI: A Pipeline-Wide Evaluation of Multi-Stage Mitigation Strategies

Shane Kennedy[2], Michael Mayowa Farayola[1,*], Daniel Kelly, Irina Tal[1], Takfarinas Saber[4], Regina Connolly[3] and Malika Bendechache[5]

[1]LERO Research Centre, School of Computing, Dublin City University, Dublin, Ireland
[2]School of Computer Science, University of Galway, Galway, Ireland
[3]LERO Research Centre, School of Business, Dublin City University, Dublin, Ireland
[4]Lero Reseach Centre, School of Computer Science, University of Galway, Ireland
[5]Lero & ADAPT Research Centres, School of Computer Science, University of Galway, Ireland

Abstract

Fairness in AI systems is critical in high-stakes domains such as healthcare, where biased predictions can exacerbate existing disparities. This paper presents an empirical evaluation of a three-stage fairness pipeline, integrating pre-processing (Disparate Impact Remover), in-processing (Exponentiated Gradient Reduction), and post-processing (Equalized Odds Optimization), on a real-world healthcare dataset from Ireland. We construct an intersectional demographic attribute to audit disparities across race, gender, and age. Our results show that multi-stage fairness interventions can reduce subgroup disparities with minimal loss in predictive performance. However, integrating fairness techniques may introduce fairness and performance trade-offs. These findings highlight the importance of holistic, intersectional fairness auditing and the need for careful design of fairness-enhancing pipelines in real-world applications.

Keywords

Trustworthy AI, Algorithmic Fairness, Healthcare AI, Intersectional Bias, Multi-Stage Mitigation,

1. Introduction

Artificial Intelligence (AI) systems are increasingly used to support high-stakes decisions in healthcare [1]. As these systems shape real-world outcomes, their *trustworthiness*, particularly in terms of fairness, has become a central concern for researchers and policymakers [2]. In domains like healthcare, fairness is not just a technical goal but an ethical necessity, as biased predictions can lead to inequitable access or harm to already marginalized groups. This concern has been codified in regulatory frameworks such as the EU Artificial Intelligence Act [3] and the NIST AI Risk Management Framework [4], which identify healthcare AI as a high-risk application requiring bias monitoring, transparency, and risk mitigation throughout the AI lifecycle.

Predictive models are applied in healthcare, influencing treatment decisions, triage, and quality-of-care metrics. If such models reinforce historical disparities, they risk disproportionately affecting vulnerable populations [5]. For example, satisfaction prediction tools that ignore demographic nuance may under-detect dissatisfaction in groups like older women of color, skewing institutional metrics and misguiding interventions. Unlike fairness studies in other domains, such as criminal justice [6], healthcare settings introduce unique challenges, including clinical heterogeneity, strict privacy constraints, and ethical obligations to minimize harm. These factors make intersectional fairness particularly critical.

TRUST-AI: The European Workshop on Trustworthy AI. Organized as part of the European Conference of Artificial Intelligence - ECAI 2025. October 2025, Bologna, Italy.

*Corresponding author.

S.Kennedy58@universityofgalway.ie (S. Kennedy); michael.farayola2@mail.dcu.ie (M. M. Farayola); dnl.kelly1@gmail.com (D. Kelly); irina.tal@dcu.ie (I. Tal); takfarinas.saber@universityofgalway.ie (T. Saber); regina.connolly@dcu.ie (R. Connolly); malika.bendechache@universityofgalway.ie (M. Bendechache)

0009-0002-3495-4155 (M. M. Farayola); 0000-0001-9656-668X (I. Tal); 0000-0003-2958-7979 (T. Saber); 0000-0003-3196-2889 (R. Connolly); 0000-0003-0069-1860 (M. Bendechache)

To mitigate these risks, fairness-enhancing interventions have been developed across the Machine Learning (ML) pipeline, pre-processing (e.g., data transformation), in-processing (e.g., fairness-aware training), and post-processing (e.g., bias output corrections) [6]. However, most studies apply these interventions in isolation using benchmark datasets with limited demographic complexity. There is little empirical evidence on how multi-stage fairness pipelines perform in real-world, intersectional contexts [7]. Fairness audits also often assess single attributes like race or gender, despite the reality that individuals experience overlapping disadvantages [8, 9]. Drawing on the framework of *intersectionality* [10], we argue that fairness assessments must account for these compounding effects. A model may appear fair across race or gender individually, but still disadvantage Black women or elderly Latina patients, biases easily missed by one-dimensional evaluations.

In this study, we evaluate a three-stage fairness pipeline comprising Disparate Impact Remover (pre-processing), Exponentiated Gradient Reduction (in-processing), and Equalized Odds (post-processing). We apply it to a real-world healthcare dataset from Ireland, using a composite intersectional attribute (race, gender, and age) across ten demographic subgroups.

We hypothesize that multi-stage fairness interventions will improve fairness across subgroups without significantly degrading predictive performance. To evaluate this, we formulate the following research questions: (1) How do fairness techniques perform when applied in combination across the pipeline? (2) Do multi-stage methods reduce disparities more effectively than single-axis evaluations suggest? (3) What trade-offs arise between fairness and performance in a sensitive healthcare task?

Moreover, to ensure practical applicability, we quantify fairness-performance trade-offs, highlighting how equity gains compare to changes in predictive metrics, such as F1 score and accuracy. Our results show that integrated fairness strategies can reduce disparities with minimal loss in performance. We advocate for fairness audits that span the entire ML pipeline and incorporate intersectional analysis to support ethically robust AI systems.

2. Trustworthy AI and Fairness: Related Work

The principles of trustworthy AI, including fairness, transparency, and accountability, are foundational to policy frameworks such as the European Commission's High-Level Expert Group on AI [11]. In ML, fairness is commonly evaluated using metrics such as Statistical Parity Difference (SPD), Disparate Impact (DI), Equal Opportunity Difference (EOD), and Average Odds Difference (AOD), which often conflict (e.g., optimizing SPD may worsen EOD) and require careful trade-off management [12].

Fairness-aware ML techniques span the full model development lifecycle. Pre-processing methods, such as Disparate Impact Remover (DIR) [13], attempt to mitigate bias in input data. In-processing approaches, such as Exponentiated Gradient Reduction (EGR) [14], optimize fairness constraints during training, whereas post-processing techniques, like Equalized Odds adjustment (EO) [15], operate on model outputs. Surveys such as [16] emphasize that most studies assess these techniques in isolation using benchmark datasets, often lacking demographic nuance and real-world complexity.

A notable exception is Farayola et al. [6, 17], who propose an integrative fairness framework in recidivism prediction. Their work integrates techniques across all three stages (pre-, in-, and post-processing phases) within a multi-objective optimization setting. They demonstrate that specific combinations (e.g., DIR + EGR + EO) can enhance fairness across multiple metrics with minimal loss in accuracy. While this represents a comprehensive multi-stage approach, it is limited to recidivism prediction. It does not address healthcare-specific challenges (e.g., clinical heterogeneity, missing data). Moreover, its applicability to real-world healthcare settings remains untested.

In healthcare, fairness challenges are compounded by clinical complexity (e.g., missing data, label noise), social determinants of health, and intersectional disparities [18, 19]. Valentine et al. [19] emphasize the importance of accounting for intersecting factors, such as race, sex, and socioeconomic status, when assessing diagnostic fairness. Huang et al. [20] conduct a scoping review and find that most fair ML applications in healthcare remain limited to single-attribute assessments, with intersectionality rarely operationalized.

To address this, the technical fairness literature has proposed intersectional frameworks. Foulds et al. [21] formalize fairness from an intersectional perspective by incorporating subgroup-level constraints into model objectives. Kearns et al. [10] propose auditing and learning algorithms that ensure fairness across a rich space of subgroups, aiming to prevent fairness gerrymandering. However, these algorithms are largely evaluated on synthetic or benchmark datasets.

Recent studies have begun to bridge this gap. Ramachandranpillai et al. [22] investigate intersectional bias mitigation in multimodal clinical prediction using the MIMIC-IV (Medical Information Mart for Intensive Care IV) dataset, which contains de-identified clinical data from patients in intensive care units (ICUs) and emergency departments (EDs). They demonstrate how biases vary across demographic intersections and data modalities, highlighting the importance of tailored fairness interventions in complex clinical settings. Wang and Yang [23] propose FairGrad, which aligns gradient updates with subgroup fairness objectives in sepsis prediction; however, their evaluation is limited to a single clinical task. While these works represent progress, they often apply to narrow clinical use cases and do not fully evaluate the cumulative interaction of fairness interventions across the ML pipeline.

Three key gaps persist: (1) Few studies evaluate end-to-end fairness pipelines (pre-, in-, post-processing) in real-world healthcare, (2) Scalable intersectional audits are lacking, particularly for overlapping subgroups (e.g., race $\times$ gender $\times$ age), (3) Trade-off quantification across pipeline stages and subgroups remains understudied.

Our work addresses these gaps by evaluating one of the three-stage integrated fairness-enhancing mitigation models identified in [6], namely DIR, EGR, and EO, on a real-world clinical dataset. We define and audit ten intersectional subgroups (race $\times$ gender $\times$ age), quantifying cumulative fairness effects and subgroup-specific performance trade-offs. Building on the integrative approach of Farayola et al. [6], our study uniquely applies it to healthcare, providing a more comprehensive and ethically grounded approach to trustworthy AI in practice.

3. Data and Ethical Considerations

This study utilizes a sensitive, real-world dataset from a healthcare organization in Ireland, containing detailed member-level information gathered through surveys, operational records, and structured metadata. The dataset is extensive, comprising 376,357 member records and 177 features. The objective is to predict individuals likely to become "detractors", members who express negative satisfaction feedback, which serves as a critical indicator of perceived quality in insurance coverage.

The dataset includes both numerical and categorical features, including sensitive attributes such as self-reported race/ethnicity, gender, and age. These were combined into a composite attribute, RACE_GENDER_AGE, to enable intersectional fairness analysis across ten distinct subgroups (e.g., White Male 65+, Latino Female 65+, White Female <65). Rare combinations were grouped into an "Other" category. The resulting minimum number of records per subgroup was 9,217 ensuring each subgroup had a credible volume of data.

To mitigate proxy discrimination, we excluded features highly correlated with protected or socioeconomic attributes, such as regional census indicators and prior-year quality scores. Categorical variables were one-hot encoded, while missing values were imputed using the mean (numerical) or mode (categorical).

Privacy and ethical safeguards were strictly followed. All identifiers were anonymized, and variable names were withheld following the healthcare provider's policies and all relevant data protection regulations. The dataset cannot be publicly released, and all analyses were conducted internally following an ethical review. Hence, this dataset enables a rare evaluation of fairness-enhancing techniques in a complex, real-world setting, where demographic nuance, imperfect data, and strict privacy constraints mirror the practical challenges of AI deployment far more closely than public benchmarks.

4. Methodology

We evaluate a three-stage fairness pipeline comprising pre-processing (Disparate Impact Remover), in-processing (Exponentiated Gradient Reduction with XGBoost), and post-processing (Equalized Odds). Performance and fairness metrics are reported at the intersectional subgroup level.

4.1. Data Preparation and Splitting

The dataset was filtered to exclude records with missing target labels or demographic attributes. Features strongly correlated with protected variables were removed to mitigate proxy bias. Categorical variables were one-hot encoded, and missing values were imputed (mean for numeric, mode for categorical).

To facilitate intersectional fairness analysis, we constructed a composite attribute (RACE_GENDER_AGE) that combines race, gender, and age, resulting in ten subgroups. The privileged group ("White Male 65+") was chosen based on domain knowledge and observed outcome advantages in the dataset, with all others treated as unprivileged.

The data was partitioned into three subsets: 50% was used to train the Base XGBoost model, the Disparate Impact Remover (DIR), the Exponentiated Gradient Reduction (EGR), and the combined DIR + EGR models; 40% was reserved for testing these models and fitting the Equalized Odds (EO) post-processing algorithm using their predictions; and the remaining 10% was held out for the final performance and fairness evaluation of the models with EO post-processing applied.

4.2. Fairness Techniques and Evaluation Metrics

We implement a three-stage fairness pipeline comprising pre-processing, in-processing, and post-processing steps. In the pre-processing stage, the Disparate Impact Remover (DIR) [13] modifies feature distributions to reduce dependence on protected attributes. For in-processing, Exponentiated Gradient Reduction (EGR) [14] incorporates fairness constraints during model training; in our implementation, protected attributes are omitted to improve fairness performance. We use an XGBoost classifier as the EGR estimator. Post-processing utilizes Equalized Odds (EO) [15], which applies group-specific adjustments to equalize false positive and false negative rates using predictions from the EGR model. Classification thresholds were tuned per demographic group by maximizing F1 scores over a grid of values (0.10–0.89).

We report classification metrics including accuracy, balanced accuracy, F1 score, recall, and AUC-ROC, disaggregated by intersectional subgroup. Fairness is evaluated using four group metrics: Disparate Impact (DI), Statistical Parity Difference (SPD), Equal Opportunity Difference (EOD), and Average Odds Difference (AOD), reflecting disparities in outcome rates and error rates between privileged and unprivileged groups. All fairness metrics were computed using the AIF360 library [24].

4.3. Fairness Results and Analysis

We evaluated each pipeline configuration across ten intersectional subgroups defined by race, gender, and age, using Disparate Impact (DI), Statistical Parity Difference (SPD), Equal Opportunity Difference (EOD), and Average Odds Difference (AOD), with ideal values of 1.0, 0.0, 0.0, and 0.0, respectively, see Table 1.

The base model, without fairness interventions, exhibited notable disparities. For example, *Asian Male Age 65+* recorded DI = 0.33, SPD = -0.57, EOD = -0.55, and AOD = -0.54, highlighting significant disadvantage and the need for intervention.

Applying DIR as a pre-processing step improved DI for some subgroups (e.g., to 0.40 for *Asian Male Age 65+*), yet SPD, EOD, and AOD remained unfavourable. This suggests that while distributional correction can partially alleviate disparity in outcome rates, it is insufficient on its own for deeper model biases.

EGR, the in-processing method, delivered broader fairness gains. All groups attained DI values between 0.99 and 1.17, and other disparity metrics stayed between -0.01 and 0.15. Notably, *Latino Female*

Age 65+ reached DI = 1.02 and SPD = 0.01. Importantly, *Asian Male Age 65+* with (DI = 0.34) in the base model, improved to 0.99 with EGR. On average, the F1 score remained 93% of the base result.

EO post-processing was not effective in reducing EOD and AOD, and its ability to improve DI was mixed. For example, *Asian Male Age 65+* still showed DI = 0.34, reflecting that post-hoc corrections are more effective when integrated with earlier-stage interventions.

The DIR+EGR configuration demonstrated strong synergy, with many subgroups reaching near-ideal fairness levels (e.g., DI between 0.79 and 0.98 and SPD between -0.01 and -0.16). On average, the F1 score remained 94% of the base result. This balance suggests that integrating fairness-enhancing techniques can improve fairness without a significant reduction in performance.

The EGR+EO pipeline offered the most consistent improvements across fairness metrics while minimizing performance trade-offs. For instance, *Latino Male Age 65+* achieved DI = 0.99, SPD = 0.0, EOD = 0.01, and AOD = 0.02, demonstrating effective fairness alignment with retained model performance.

In contrast, DIR+EO (without in-processing) produced less stable outcomes. Fairness metrics did not materially improve and performance metrics worsened for most groups (e.g., F1 score reduced from 0.26 to 0.16 for *Black Male Age 65+*). This underscores the critical role of model-level adjustments in fairness optimization.

The full pipeline (DIR + EGR + EO) achieves a significant impact, but fairness outcomes were not consistently improved. For instance, the DI for *Asian Male Age 65+* remained below parity at 0.78, and subgroup-level F1 score declined in cases such as *Black Male 65+*. These results indicate that multi-stage mitigation must be applied with careful calibration to avoid overcorrection and performance degradation.

Integrating fairness interventions can reduce disparities, though often at the cost of performance metrics such as the F1 score if not carefully configured. EGR+EO improved AOD by 32% on average across sub-groups with 93% F1 retention, and DIR+EGR achieved near-parity DI with 94% F1 retention. In contrast, configurations like DIR+EO or DIR+EGR+EO risk fairness drift or lower F1 scores. These findings highlight that no single mitigation technique performs best across all subgroups or metrics. Effectiveness depends on data characteristics, underscoring the need for tailored combinations that address specific disparities more effectively.

Table 1: Performance and fairness metrics across intersectional subgroups for each pipeline configuration, with negative values shown in braces.

Model	Description	Best Threshold	Accuracy	F1-Score	Recall	AUC	DI	SPD	EOD	AOD
Base	White, Female, Age 65+	0.56	0.72	0.26	0.46	0.66	0.99	(0.01)	(0.01)	(0.01)
	Black, Male, Age 65+	0.57	0.74	0.26	0.44	0.67	0.98	(0.02)	(0.03)	-
	Latino, Male, Age 65+	0.59	0.76	0.28	0.42	0.67	0.85	(0.12)	(0.12)	(0.11)
	White, Male, Age under 65	0.56	0.73	0.27	0.48	0.67	0.78	(0.18)	(0.18)	(0.17)
	Latino, Female, Age 65+	0.60	0.79	0.28	0.38	0.68	0.80	(0.17)	(0.16)	(0.17)
	White, Female, Age under 65	0.56	0.73	0.27	0.48	0.67	0.72	(0.23)	(0.23)	(0.24)
	Black, Female, Age 65+	0.58	0.75	0.26	0.41	0.66	0.78	(0.19)	(0.18)	(0.20)
	Asian, Male, Age 65+	0.58	0.75	0.27	0.46	0.66	0.33	(0.57)	(0.55)	(0.54)
	Other	0.57	0.74	0.27	0.47	0.68	0.44	(0.48)	(0.47)	(0.47)
DIR	White, Female, Age 65+	0.55	0.73	0.25	0.43	0.66	1.00	-	-	(0.01)
	Black, Male, Age 65+	0.54	0.72	0.27	0.48	0.67	0.96	(0.03)	(0.04)	(0.02)
	Latino, Male, Age 65+	0.52	0.68	0.27	0.53	0.66	0.83	(0.14)	(0.14)	(0.13)
	White, Male, Age under 65	0.56	0.74	0.27	0.44	0.67	0.73	(0.23)	(0.22)	(0.22)
	Latino, Female, Age 65+	0.56	0.76	0.26	0.42	0.66	0.74	(0.22)	(0.21)	(0.22)
	White, Female, Age under 65	0.58	0.77	0.28	0.39	0.67	0.66	(0.29)	(0.28)	(0.30)
	Black, Female, Age 65+	0.58	0.78	0.27	0.39	0.67	0.70	(0.25)	(0.24)	(0.27)
	Asian, Male, Age 65+	0.59	0.80	0.27	0.35	0.67	0.40	(0.51)	(0.49)	(0.48)
	Other	0.58	0.78	0.27	0.38	0.66	0.43	(0.48)	(0.47)	(0.48)
EGR	White, Female, Age 65+	0.59	0.65	0.25	0.53	0.63	1.01	0.01	0.01	-
	Black, Male, Age 65+	0.67	0.66	0.24	0.50	0.63	1.07	0.05	0.05	0.06
	Latino, Male, Age 65+	0.83	0.72	0.25	0.44	0.63	1.02	0.02	0.01	0.04
	White, Male, Age under 65	0.72	0.67	0.25	0.50	0.63	1.17	0.11	0.11	0.15
	Latino, Female, Age 65+	0.38	0.60	0.24	0.61	0.63	1.02	0.01	0.02	0.02
	White, Female, Age under 65	0.38	0.61	0.28	0.62	0.64	1.15	0.10	0.10	0.12
	Black, Female, Age 65+	0.87	0.79	0.26	0.34	0.65	1.00	-	-	0.01
	Asian, Male, Age 65+	0.63	0.66	0.25	0.49	0.61	0.99	(0.01)	-	0.04
	Other	0.83	0.71	0.24	0.44	0.63	1.05	0.03	0.04	0.05

EO	White, Female, Age 65+	-	0.83	0.20	0.24	-	1.00	-	(0.01)	-
	Black, Male, Age 65+	-	0.82	0.20	0.24	-	0.99	-	(0.01)	0.01
	Latino, Male, Age 65+	-	0.71	0.22	0.34	-	0.81	(0.16)	(0.16)	(0.14)
	White, Male, Age under 65	-	0.71	0.26	0.41	-	0.76	(0.20)	(0.19)	(0.21)
	Latino, Female, Age 65+	-	0.69	0.25	0.37	-	0.77	(0.19)	(0.18)	(0.18)
	White, Female, Age under 65	-	0.68	0.27	0.45	-	0.71	(0.24)	(0.23)	(0.25)
	Black, Female, Age 65+	-	0.71	0.27	0.44	-	0.76	(0.20)	(0.20)	(0.20)
	Asian, Male, Age 65+	-	0.51	0.32	0.66	-	0.34	(0.55)	(0.53)	(0.53)
	Other	-	0.55	0.28	0.59	-	0.44	(0.47)	(0.46)	(0.46)
DIR+EGR	White, Female, Age 65+	0.81	0.71	0.24	0.44	0.63	0.98	(0.01)	(0.01)	(0.02)
	Black, Male, Age 65+	0.35	0.67	0.24	0.50	0.62	0.96	(0.03)	(0.03)	(0.01)
	Latino, Male, Age 65+	0.31	0.65	0.24	0.50	0.62	0.95	(0.03)	(0.03)	(0.01)
	White, Male, Age under 65	0.63	0.68	0.26	0.53	0.64	0.87	(0.09)	(0.09)	(0.07)
	Latino, Female, Age 65+	0.33	0.67	0.27	0.56	0.66	0.92	(0.06)	(0.05)	(0.05)
	White, Female, Age under 65	0.85	0.76	0.27	0.39	0.64	0.85	(0.11)	(0.10)	(0.10)
	Black, Female, Age 65+	0.81	0.71	0.26	0.46	0.64	0.84	(0.12)	(0.11)	(0.11)
	Asian, Male, Age 65+	0.31	0.66	0.25	0.52	0.62	0.79	(0.16)	(0.15)	(0.12)
	Other	0.64	0.68	0.25	0.49	0.62	0.81	(0.14)	(0.13)	(0.13)
EGR+EO	White, Female, Age 65+	-	0.69	0.23	0.51	-	1.01	0.01	0.01	(0.01)
	Black, Male, Age 65+	-	0.69	0.19	0.38	-	1.05	0.03	0.03	0.06
	Latino, Male, Age 65+	-	0.66	0.24	0.44	-	0.99	-	(0.01)	0.02
	White, Male, Age under 65	-	0.73	0.24	0.32	-	1.15	0.10	0.10	0.13
	Latino, Female, Age 65+	-	0.67	0.30	0.47	-	1.00	-	0.01	0.01
	White, Female, Age under 65	-	0.75	0.23	0.38	-	1.14	0.10	0.10	0.10
	Black, Female, Age 65+	-	0.66	0.28	0.49	-	0.98	(0.01)	-	-
	Asian, Male, Age 65+	-	0.62	0.28	0.38	-	0.99	-	(0.01)	0.05
	Other	-	0.68	0.31	0.46	-	1.03	0.02	0.03	0.03
DIR+EO	White, Female, Age 65+	-	0.84	0.20	0.21	-	1.00	-	-	(0.02)
	Black, Male, Age 65+	-	0.81	0.16	0.20	-	0.95	(0.04)	(0.05)	(0.02)
	Latino, Male, Age 65+	-	0.74	0.21	0.29	-	0.81	(0.16)	(0.16)	(0.15)
	White, Male, Age under 65	-	0.71	0.23	0.33	-	0.76	(0.20)	(0.20)	(0.19)
	Latino, Female, Age 65+	-	0.72	0.29	0.39	-	0.78	(0.19)	(0.17)	(0.20)
	White, Female, Age under 65	-	0.68	0.22	0.40	-	0.71	(0.24)	(0.24)	(0.27)
	Black, Female, Age 65+	-	0.72	0.21	0.34	-	0.76	(0.21)	(0.20)	(0.22)
	Asian, Male, Age 65+	-	0.54	0.34	0.55	-	0.40	(0.51)	(0.50)	(0.49)
	Other	-	0.56	0.27	0.52	-	0.46	(0.47)	(0.46)	(0.46)
DIR+EGR+EO	White, Female, Age 65+	-	0.73	0.23	0.46	-	1.00	-	-	(0.02)
	Black, Male, Age 65+	-	0.69	0.18	0.36	-	0.96	(0.03)	(0.03)	0.01
	Latino, Male, Age 65+	-	0.68	0.26	0.43	-	0.93	(0.05)	(0.04)	(0.03)
	White, Male, Age under 65	-	0.66	0.26	0.48	-	0.88	(0.09)	(0.08)	(0.08)
	Latino, Female, Age 65+	-	0.68	0.27	0.44	-	0.93	(0.05)	(0.04)	(0.04)
	White, Female, Age under 65	-	0.67	0.25	0.54	-	0.89	(0.08)	(0.07)	(0.09)
	Black, Female, Age 65+	-	0.66	0.27	0.53	-	0.86	(0.11)	(0.10)	(0.11)
	Asian, Male, Age 65+	-	0.60	0.35	0.55	-	0.78	(0.16)	(0.15)	(0.14)
	Other	-	0.62	0.30	0.52	-	0.82	(0.13)	(0.12)	(0.12)

4.4. Key Insights from the Analysis

The empirical evaluation reveals several critical insights about the interplay between fairness interventions and predictive performance across intersectional subgroups, see Table 1.

Insight 1: Multi-stage interventions outperform single-stage ones. Our results confirm that fairness is most effectively improved when mitigation techniques are applied at multiple stages of the machine learning pipeline. Configurations such as *DIR + EGR* and *EGR + EO* consistently delivered more equitable outcomes across metrics like DI, SPD, and AOD, compared to any individual intervention alone. These combinations reduced disparities while preserving substantial performance, supporting the notion that fairness must be embedded throughout the pipeline rather than treated in isolation.

Insight 2: Trade-offs between fairness and performance are non-uniform. Some intervention strategies improved fairness with minimal performance impact, while others led to subgroup-specific degradation. For instance, *EGR+EO* achieved low disparity scores alongside strong recall and F1 performance. By contrast, *DIR+EGR+EO* produced modest gains in recall for several groups but was less stable overall, with some subgroups (e.g., *White Male under 65*) experiencing notable performance declines. These outcomes indicate that fairness-utility trade-offs vary by configuration and cannot be generalized.

Insight 3: High Fairness Scores Can Mask Instability. While most DI values clustered around 1.0 under well-calibrated pipelines, some subgroups showed exaggerated improvements. For exam-

ple, *Latino Male Age 65+* reached DI = 0.93 under *DIR + EO*, and 0.82 under *DIR+EGR+EO*, indicating performance shifts not necessarily aligned with improved fairness. These patterns may reflect overcompensation or instability, suggesting the need for holistic subgroup-level audits that go beyond average metrics to ensure fair and balanced outcomes.

Insight 4: Intersectional subgroup evaluation is crucial. Certain groups, particularly *Asian Male Age 65+*, consistently underperformed in both fairness and performance metrics, even under fairness-aware models. For example, this group maintained a DI below 0.5 across several configurations. This illustrates how single-axis assessments (e.g., race or gender alone) can obscure compounded disadvantage. Intersectional subgroup analysis is essential to reveal complex and persistent inequities that aggregated evaluations would miss.

Insight 5: Post-processing is most effective when preceded by upstream corrections. Equalized Odds (EO) post-processing improved fairness metrics such as EOD and AOD, particularly when combined with fairness-aware training (e.g., EGR). Applied alone, EO delivered modest fairness gains but could not overcome entrenched disparities from biased inputs or model structures. This underscores EO's role as a valuable final-stage tool, but only when upstream bias has been addressed through preprocessing or in-processing.

These insights highlight the need for fairness-aware model development that considers both technical and social contexts. They underscore the importance of testing interventions in combination and across diverse subgroups, as effects are often variable and configuration-dependent.

5. Discussion

Our results show that fairness-enhancing interventions applied across the ML pipeline can improve equity across intersectional subgroups with minimal compromise to overall predictive performance. The most balanced configuration, combining Exponentiated Gradient Reduction (EGR) with Equalized Odds (EO), consistently reduced disparities across multiple fairness metrics while preserving performance metrics such as F1 score. This supports the effectiveness of aligning fairness constraints during training with calibrated post-hoc adjustments.

Nonetheless, fairness–utility trade-offs remain nuanced and context-dependent. For instance, while the *DIR+EGR+EO* configuration improved F1 in some subgroups, it also introduced uneven fairness outcomes. *Latino Male Age 65+* had DI = 0.93 under this configuration, suggesting under-correction rather than overcompensation. These outcomes highlight the importance of carefully tuned fairness constraints and subgroup-aware threshold optimization, especially when working with smaller or structurally marginalized populations.

The added complexity of multi-stage fairness pipelines also raises interpretability and transparency concerns. Although removing protected attributes during training helps limit discriminatory influence, it may hinder the model's capacity to detect and address embedded inequities. Future research should explore how causal inference and explainability techniques can bridge this gap and improve accountability in fairness-aware systems.

Although the proprietary nature of our dataset limits full reproducibility, it reflects common challenges in real-world healthcare settings, including privacy constraints and complex demographics. Despite this limitation, our findings support the hypothesis that multi-stage fairness interventions can promote equitable outcomes with manageable performance trade-offs (RQ1, RQ3). Furthermore, the intersectional evaluation framework uncovered disparities that single-axis analyses would have missed (RQ2), reinforcing the value of comprehensive fairness auditing in high-stakes domains.

5.1. Limitations

This study has several limitations. First, while we evaluated multiple fairness metrics, including Disparate Impact, Statistical Parity Difference, Equal Opportunity Difference, and Average Odds Difference, our primary modeling focus did not incorporate fairness constraints like Equalized Odds during training or calibration, which may yield different trade-offs. Second, reliance on proprietary healthcare data

limits reproducibility and precludes external benchmarking. Third, the fairness–performance dynamics observed in this specific healthcare context may not generalize to other domains or geographic settings. Fourth, Equalized Odds (EO) post-processing assumes access to true outcome labels at deployment time—a strong requirement that may limit its real-world applicability. Additionally, EO models were evaluated on a separate held-out test set (10%) following calibration on a distinct test split (40%) using predictions from upstream models. While this setup is required by the EO algorithm, which learns group-specific adjustments from predicted and true labels, it limits direct comparability with other models evaluated only on the 40% test set and may introduce confounding effects due to distributional differences. Lastly, some intersectional subgroups were underrepresented, potentially leading to unstable metric estimates and wide variance. Future work should explore fairness auditing techniques that are robust to low-sample settings and demographic imbalance.

6. Conclusion and Future Work

This paper presented an empirical evaluation of a multi-stage fairness pipeline applied to a real-world healthcare dataset, integrating Disparate Impact Remover (DIR), Exponentiated Gradient Reduction (EGR), and Equalized Odds Optimization (EO). Our results demonstrate that when strategically integrated, fairness interventions can reduce disparities across intersectional subgroups while maintaining acceptable predictive performance.

Our work provides an actionable framework for compliance with emerging regulations like the EU AI Act [3] and NIST AI RMF [4], which mandate bias mitigation in high-risk AI systems. By integrating intersectional audits, we address their call for transparency and fairness across demographic subgroups.

The combination of EGR and EO proved most balanced, lowering fairness metrics such as AOD and EOD with minimal accuracy loss. However, certain groups, most notably Asian Male Age 65+, continued to experience inequities, underscoring the limitations of current techniques and the need for subgroup-sensitive approaches.

EO was most effective when applied after upstream mitigation, reinforcing its role as a complementary rather than standalone tool. While DIR+EGR+EO improved performance for some groups, it introduced instability, highlighting the need for calibrated design.

Future work will integrate fairness constraints into training, apply causal inference to structural disparities, and expand to other domains. We also plan to apply explainability tools to enhance stakeholder trust and explore subgroup-sensitive fairness interventions. Overall, our findings support the value of multi-stage, intersectionally-aware fairness pipelines as a foundation for responsible and trustworthy AI in healthcare.

Acknowledgments

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

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] S. Montani, M. Striani, Artificial intelligence in clinical decision support: a focused literature survey, Yearbook of medical informatics 28 (2019) 120–127.

[2] M. M. Farayola, I. Tal, R. Connolly, T. Saber, M. Bendechache, Ethics and trustworthiness of ai for predicting the risk of recidivism: A systematic literature review, Information 14 (2023) 426.
[3] European Parliament and Council, Regulation (EU) 2024/1689 of the European Parliament and of the Council, Official Journal of the European Union, L 135, 1–121, 2024. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj.
[4] N. AI, Artificial intelligence risk management framework (ai rmf 1.0) (2023) 100–1. doi:https://doi.org/10.6028/NIST.AI.100-1.
[5] M. Liu, Y. Ning, S. Teixayavong, X. Liu, M. Mertens, Y. Shang, X. Li, D. Miao, J. Liao, J. Xu, et al., A scoping review and evidence gap analysis of clinical ai fairness, npj Digital Medicine 8 (2025) 360.
[6] M. M. Farayola, M. Bendechache, T. Saber, R. Connolly, I. Tal, Enhancing algorithmic fairness: Integrative approaches and multi-objective optimization application in recidivism models, in: Proceedings of the 19th International Conference on Availability, Reliability and Security, 2024, pp. 1–10.
[7] A. Wang, V. V. Ramaswamy, O. Russakovsky, Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation, in: Proceedings of the 2022 ACM conference on fairness, accountability, and transparency, 2022, pp. 336–349.
[8] U. Gohar, L. Cheng, A survey on intersectional fairness in machine learning: Notions, mitigation, and challenges, arXiv preprint arXiv:2305.06969 (2023).
[9] A. Ovalle, A. Subramonian, V. Gautam, G. Gee, K.-W. Chang, Factoring the matrix of domination: A critical review and reimagination of intersectionality in ai fairness, in: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, 2023, pp. 496–511.
[10] M. Kearns, S. Neel, A. Roth, Z. S. Wu, Preventing fairness gerrymandering: Auditing and learning for subgroup fairness, in: International conference on machine learning, PMLR, 2018.
[11] H. AI, High-level expert group on artificial intelligence, 2019.
[12] S. Barocas, M. Hardt, A. Narayanan, Fairness and machine learning: Limitations and opportunities, MIT press, 2023.
[13] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, S. Venkatasubramanian, Certifying and removing disparate impact, in: proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, 2015, pp. 259–268.
[14] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, H. Wallach, A reductions approach to fair classification, in: International conference on machine learning, PMLR, 2018, pp. 60–69.
[15] M. Hardt, E. Price, N. Srebro, Equality of opportunity in supervised learning, Advances in neural information processing systems 29 (2016).
[16] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Computing Surveys (CSUR) 54 (2021) 1–35.
[17] M. M. Farayola, I. Tal, T. Saber, R. Connolly, M. Bendechache, A fairness-focused approach to recidivism prediction: implications for accuracy, trust, and equity, AI & SOCIETY (2025) 1–19.
[18] B. Koçak, A. Ponsiglione, A. Stanzione, C. Bluethgen, J. Santinha, L. Ugga, M. Huisman, M. E. Klontzas, R. Cannella, R. Cuocolo, Bias in artificial intelligence for medical imaging: fundamentals, detection, avoidance, mitigation, challenges, ethics, and prospects, Diagnostic and interventional radiology 31 (2025) 75.
[19] A. A. Valentine, A. W. Charney, I. Landi, Fair machine learning for healthcare requires recognizing the intersectionality of sociodemographic factors, a case study, arXiv preprint arXiv:2407.15006 (2024).
[20] Y. Huang, J. Guo, W.-H. Chen, H.-Y. Lin, H. Tang, F. Wang, H. Xu, J. Bian, A scoping review of fair machine learning techniques when using real-world data, Journal of Biomedical Informatics (2024) 104622.
[21] J. R. Foulds, R. Islam, K. N. Keya, S. Pan, An intersectional definition of fairness, in: 2020 IEEE 36th International Conference on Data Engineering (ICDE), IEEE, 2020, pp. 1918–1921.
[22] R. Ramachandranpillai, K. Sampath, A. Mohammad, M. Alikhani, Fairness at every intersection: Uncovering and mitigating intersectional biases in multimodal clinical predictions, arXiv preprint arXiv:2412.00606 (2024).

[23] X. Wang, C. C. Yang, Enhancing multi-attribute fairness in healthcare predictive modeling, arXiv preprint arXiv:2501.13219 (2025).
[24] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, et al., Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias, IBM Journal of Research and Development 63 (2019) 4–1.