ECAI 2025
I. Lynce et al. (Eds.)

doi:10.3233/FAIA251105

Achieving Intersectional Algorithmic Fairness by Constructing a Maximal Correlation Latent Space

Luca Giuliani[a,*] **and Michele Lombardi**[a]

[a]University of Bologna, Department of Computer Science and Engineering (DISI)

Abstract. Recent developments in algorithmic fairness started to investigate the interaction between multiple sensitive information through an intersectional perspective. We introduce a new definition of intersectional fairness based on a multivariate extension of the Generalized Disparate Impact (GeDI). Our approach leverages a neural network to transform multiple protected groups into a univariate latent space that maximizes correlation with the target, effectively capturing unfairness across all potential subgroups even with limited data samples. Empirical evaluations on several benchmarks demonstrate that our method can be effectively used as a loss regularizer during neural network training, offering stronger performance guarantees compared to existing intersectional statistical parity definitions while also allowing to manage continuous inputs and targets.

1 Introduction

The influence of Machine Learning (ML) and Artificial Intelligence (AI) on society has risen significantly in recent years. As data-driven technologies become more prominent, concerns about their behavior arise, followed by technical guidelines [18] and legislative actions [25]. In such documents, a major emphasis is reserved to the trustworthiness of these systems, among which discrimination prevention is one of the key points. Public opinion awareness about racial and gender biases in AI-based support decision systems was first raised by the ProPublica study of the COMPAS criminal justice software [2], but many other examples can be found in hiring [9], education [21], and healthcare [33].

Analyzing algorithmic fairness is a complex task, with traditional methods like "fairness through unawareness" proven ineffective [8]. Yet, several strategies and fairness indicators have been developed in the past decade to align Machine Learning models with fairness principles, preventing discrimination against minorities [27]. Most early statistical fairness research focused on a single "protected attribute", e.g., race or gender. This setting, called independent group fairness [35], is strongly limited with respect to intersectional group fairness, a framework based on Crenshaw's theory of intersectionality [6] which aims to address stratified forms of unfairness across marginalized subgroups by taking into accounts the joint effects of multiple protected attributes at the same time. Buolamwini and Gebru [5] were the first to show that such kind of biases are already present in AI-based systems, revealing that the three main face recognition software available on the market reported strong accuracy gaps between white males, white females, black males, and black females – who were in fact the most underrepresented and discriminated subgroup. Nonetheless, building intersectionally-fair predictors poses many technical challenges, such as determining the appropriate granularity that could correctly safeguard even the sub-communities that are underrepresented in the training dataset [15].

1.1 Motivation

All the existing body of research on intersectional algorithmic fairness currently resorts to brute force enumeration techniques to quantify the opportunity gap between privileged and marginalized groups [35]. For example, in their seminal work [20], Kearns et al. introduce the notion of Subgroup Fairness, a concept that extends the definition of group fairness by accounting for intersectional, independent, and other identifiable subgroups delineated by multiple protected attributes. The intention behind this is to prevent what the authors call "Fairness Gerrymandering", i.e., the practice of correctly applying fairness to each individual group while dramatically violating such requirement on certain structured subgroups. Through this approach, statistical and individual fairness can be arbitrarily integrated, thereby combining evaluations in between personal identities and social categories defined across overlapping (sub-)communities.

Nonetheless, such enumeration-based strategy encounters a major statistical and computational barrier, since the inclusion of a vast quantity of subgroups may restrict the count of non-empty subsets, particularly in situations where limited data is available. In order to address this, Kearns et al. employ a parameter that allows to disregard a group if its representation share within the dataset is minimal, hence controlling the portion of the population that is allowed to be overlooked. This poses a significant issue, particularly considering that social communities with minimal representation are often the most marginalized and discriminated ones. On the contrary, [12] discards this bias-inducing parameter and equally weights discrimination across all identifiable subgroups in their Differentiable Fairness indicator, showing how this allows for the prevention of fairness gerrymandering effects directly from the intersectional level, hence reducing the number of subsets that need to be considered. Despite this improvement, Differential Fairness still relies on enumeration, tying it to non-empty subgroups only – albeit in a much more coarse-grained scale. Similar challenges are noted in [17] and [35], which base their fairness criteria on Equalized Odds rather than Statistical Parity, revealing that the application of subgroup fairness measures is currently limited to computationally-identifiable subgroups.

As discussed in [23], this behavior inadvertently risks the enforcement of a restrictive boundary on subgroups evaluation which would fail to protect minorities that do not fit these criteria, hence perpetuating the same fairness gerrymandering that they aim to resolve. There-

* Corresponding Author. Email: luca.giuliani13@unibo.it.

fore, motivated by this recognized issue, we propose a novel intersectional fairness indicator that does not rely on enumeration. Instead, intercepting the advancements from another branch of algorithmic fairness focused on handling continuous protected attributes rather than categorical ones [14, 16, 19, 22, 26], we assess (un)fairness by mapping the protected input features into a univariate latent space that maximizes its correlation with the output target. To the best of our knowledge, this is the first proposal of an intersectional method that is based on function approximation, enabling us to avoid enumeration issues thanks to interpolation and extrapolation abilities.

1.2 Contribution

We introduce a new definition of intersectional fairness based on a multivariate extension of the Generalized Disparate Impact (GeDI) indicator, which we previously introduced in [14]. To address the challenge of managing multiple input variables simultaneously, we adapt the semantics of our original indicator and implement a computational procedure inspired by [16]. In fact, our mapping function is a neural network trained to maximize the correlation, measured in terms of *disparate impact*, between the latent space projection and the target vector. This approach allows to achieve two key objectives:

1. it enhances the handling of underrepresented social communities by aggregating information from similar subgroups thanks to an approximation function, which improves constraint satisfaction in both train and test data;
2. it inherently deals with continuous protected attributes like age or income, as well as multi-class or continuous targets, which are currently intractable with any existing indicators.

We underline that our use of the term "intersectionality" strictly concerns its statistical implications. In that, we align with the limitations highlighted by [7] and [34], remarking that our contribution restricts to the subject of fairness measurement and enforcement, with no direct claims on how automated intersectional discrimination is affecting and will affect both individuals and society as a whole. We are aware that fairness is not a merely technical concept, and on the contrary it encompasses legal, economical, philosophical, and sociopolitical aspects, making any quantitative analysis, particularly in the intersectional domain, intrinsically limited in scope as it can only evaluate the impacts of systemic oppression without fully comprehending its underlying causes [3, 23]. Nevertheless, we chose to retain the term "intersectionality" not only due to its increasing prevalence in the literature, but also for its capacity to incentivize a deeper reflection on the effects of machine-based discrimination through the academically and culturally relevant works on the subject.

The rest of the paper is structured as follows. Section 2 surveys the state of the art concerning algorithmic fairness, with an emphasis on intersectional fairness and fairness involving continuous protected attributes. In Section 3, we explore the technicalities of our proposed indicator and its implementation, demonstrating that it maintains significant theoretical guarantees regarding its generalization abilities. Section 4 presents the outcomes of two experiments, where we empirically assess the validity of such properties across four benchmark datasets for fair machine learning, and also prove that our metric enhances constraint satisfaction compared to alternative definitions. Lastly, our concluding remarks are outlined in Section 5.

2 Background and State of the Art

Long before the rise of automated systems, human decision-makers were already engaging in discriminatory practices especially in sectors like insurance and banking. A prominent example of that is "redlining" [24], which consists in the denial of a service to marginalized communities based on their residential areas, but there exist also other forms of indirect discrimination, such as "discrimination on redundant encodings", "self-fulfilling prophecy", and "reverse tokenism". As highlighted in [8], the mere omission of the protected attributes from the training data – also defined as "fairness through unawareness" – is ineffective in preventing automated discrimination, as machine learning models often learn and perpetuate existing biases using proxy attributes that are closely correlated with the protected ones. Consequently, several indicators have been devised to assess the impact of decisions on protected groups, a concept legally referred to as *disparate impact* [10]. This term gained prominence in the U.S. legal system following the *Griggs v. Duke Power Co.* case[1], where a company was accused of systematic rejection of black candidates based on the results of intelligence tests and the absence of a high school diploma.

2.1 Independent Group Fairness

Given a classifier $\mathcal{C}$, disparate impact is mathematically translated as the degree of statistical independence between the protected attribute x and the predicted outcome $\hat{y} = \mathcal{C}(x)$. Early fairness measures like Demographic Parity (DP) focus on binary protected attributes $x \in \{m, p\}$ and binary predictions $\hat{y} \in \{+, -\}$, respectively representing the division between marginalized and privileged communities, and between positive and negative outcomes. Unfairness is then quantified as the opportunity gap between the two groups, i.e.:

$$\mathrm{DP}(x, \hat{y}) = P(\hat{y} = + \mid x = p) - P(\hat{y} = + \mid x = m) \quad (1)$$

where the probability P is empirically estimated as the ratio of the positive labels in each group. A notable extension for multi-class inputs and continuous outputs is proposed in [1], where the authors define the Disparate Impact Discrimination Index (DIDI) as the sum of the opportunity gap over each protected group with respect to the total average. Said $\mathcal{G}$ the set of protected groups and $\mathbb{E}$ the expectation over the continuous target distribution, we have:

$$\mathrm{DIDI}(x, \hat{y}) = \sum_{g \in \mathcal{G}} |\mathbb{E}[\hat{y} \mid x = g] - \mathbb{E}[\hat{y}]| \quad (2)$$

While DP and DIDI focus on the correlation between the predictions $\hat{y}$ and the protected attribute x, other measures such as Equalized Odds (EO) also take into account the ground truth vector y, aimed at measuring the degree of *conditional* independence between $\hat{y}$ and x with respect to y:

$$\begin{aligned} \mathrm{EO}(x, y, \hat{y}) = & P(\hat{y} = + \mid y = +, x = p) - \\ & P(\hat{y} = + \mid y = +, x = m) \end{aligned} \quad (3)$$

In practice, EO assesses fairness based on the gap in predictive accuracy rather than the gap in the outcome itself; for this reason, it contrasts with the legal and philosophical definition of disparate impact and aligns instead with the principle of infra-marginality. [32] defines infra-marginality as a perspective for which the differences between privileged and marginalized groups outcomes is consequential to an inherent "merit" or "risk" attribute tied to their individuals, hence any mismatch in predicted results should not be considered unfair unless related to a prediction error; on the contrary, intersectional feminists and activists stress that such differences are the consequences of systematic oppression of unprivileged groups across

[1] https://supreme.justia.com/cases/federal/us/401/424/

multiple discrimination forms such as sexism, racism, ableism, classism, homo/bi/transphobia, etc. Practically, while infra-marginality principles assume society to be fair by default, intersectional theory implies the existence of systems of oppression that limit one's potential due to higher exposition to, e.g., poverty, discrimination, and lack of access to education, resulting in both direct and indirect forms of exclusion explaining the distribution mismatch. For this reason, although acknowledging the relevance of Equalized Odds and similar metrics in several tasks like face recognition [5], we will focus on disparate impact measures for the rest of the paper as we believe that they better suit our research objectives.

Finally, on the opposite end of group fairness there is individual fairness, which is concerned in ensuring that similar individuals are granted similar opportunities. Typically, individual fairness indicators either use distance measures or some kind of counterfactual reasoning to ensure that similar outcomes are obtained within certain distance bounds in the input space. This is often in contrast with the objectives of group fairness, particularly if defined in terms of disparate impact, therefore we will not explore the nuances of individual fairness and refer interested readers to [4].

2.2 Fairness With Continuous Protected Attributes

All the indicators discussed above can only handle categorical protected attributes because of two main reasons: first, fairness definitions have historically been associated to social groups identified by demographic traits such as gender, ethnicity, sexual orientation, religious belief, and others; secondly, quantifying fairness typically requires access to probability distributions, which are considerably simpler to estimate for categorical variables as opposed to continuous ones. Consequently, numerical features such as income or age have been systematically excluded from fairness considerations.

Theoretically, continuous protected attributes can be managed through discretization; however, this method struggles in practice due to its sensitivity to bin configuration, which can be manipulated with malicious intents. As an alternative approach, [26] firstly proposed the adoption of the Hirschfeld–Gebelein–Rényi (HGR) Coefficient to measure fairness with continuous attributes, using Kernel Density Estimation to approximate necessary distributions. [16] further develops on this by employing two neural networks for the estimation of HGR, while [19] applies Kernel Density Estimation techniques directly to extend Demographic Parity to continuous attributes and [22] uses Expectation of Integral Probability to quantify the distribution mismatch between input and target. Lastly, in our previous work [14], we altered the semantics of HGR to align it with the legal definition of disparate impact, demonstrating that, for a binary protected input and a continuous target, the DIDI can be restated as:

$$\mathrm{DIDI}(x, y) = \frac{\mathrm{cov}(x, y)}{\sigma(x)^2}, \quad x \in \{0, 1\}^n, \quad y \in \mathbb{R}^n \tag{4}$$

Following, we introduced the Generalized Disparate Impact as:

$$\mathrm{GeDI}(x, y) = \max_{f \in \mathbb{R} \mapsto \mathbb{R}} \frac{\mathrm{cov}(f(x), y)}{\sigma(f(x))^2}, \quad x, y \in \mathbb{R}^n \tag{5}$$

and complemented it with a computational method for its estimation on finite datasets. The use of a mapping function f outlined in Equation (5) is derived from the theoretical definition of HGR [30], and allows to extend traditional linear correlation measures to non-linear scenarios. In this work, we adopt the same concept with the aim of capturing non-linear joint interactions among multiple protected attributes, thereby generalizing GeDI to a multivariate context.

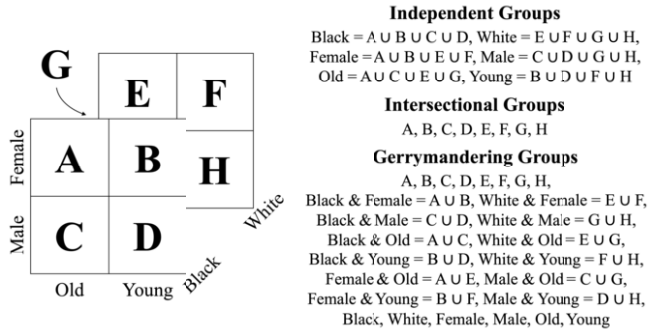


Figure 1. Independent, intersectional, and gerrymandering groups as defined by Yang et al. – the image is taken from their original paper [35].

2.3 Intersectional Fairness Definitions

Intersectional algorithmic fairness has only recently gained attention. Pioneering this field, Kearns et al. introduced in [20] the concept of Statistical Parity Subgroup Fairness (SPSF) as a way to build an intersectional definition of fairness to prevent what they called "Fairness Gerrymandering". Practically, gerrymandering refers to the different treatment of subgroups defined not only on each protected attribute $X_1, \ldots, X_p$ independently, but also on each combination of their subsets $\mathcal{S} \subseteq \{X_1, \ldots, X_p\}$, as illustrated in Figure 1. Given the respective domains $\mathcal{X}_j$ of each protected attribute j, a set of characteristic functions $g_i : \mathcal{X}_1 \times \cdots \times \mathcal{X}_p \mapsto \{0, 1\}$ is used to assign each individual to their subgroup i; then, a classifier $\mathcal{C}$ is said to be γ-fair if, for all subgroups i it holds:

$$\begin{aligned} &\alpha(X) \cdot |\beta(X)| \leq \gamma, \\ &\alpha(X) = P(g_i(X_P) = 1) \\ &\beta(X) = P(\mathcal{C}(X) = + \mid g_i(X_P) = 1) - P(\mathcal{C}(X) = +) \end{aligned} \tag{6}$$

where $X_P = \left[X_1, \ldots, X_p\right]$ is the matrix of protected features.

Foulds et al. claim that the α term does not fully respect intersectionality definitions, and accordingly adjust the semantics of SPSF into an Empirical Differential Fairness (EDF) definition [12]. Taking the same classifier $\mathcal{C}$, we can say that it is ϵ-differentially fair if:

$$e^{-\epsilon} \leq \frac{P(\mathcal{C}(X) = + \mid g_i(X_P) = 1)}{P(\mathcal{C}(X) = + \mid g_j(X_P) = 1)} \leq e^{\epsilon} \tag{7}$$

holds for every subgroup pair (i, j).

Both SPSF and EDF align with disparate impact by evaluating outcome gaps rather than predictive accuracy gaps. On the contrary, both their respective variants False Positive Subgroup Fairness and $(\epsilon_2 - \epsilon_1)$-DF Bias Amplification, together with other measures like Multi-calibration [17] and Metric-based fairness [31], extend the idea of Equalized Odds to the intersectional scenario, thus we will not include them in our further discussion or experiments for previously stated reasons. Moreover, we remark again that all these indicators rely on enumeration, thus facing technical and ethical challenges due to limited data. [15] notes that, despite inherently covering all gerrymandering groups, Differential Fairness [12] and Max-Min Fairness [13] also struggle with data sparsity, and recommends developing probabilistic methods like [28]. We claim that our approach balances both aspects, exploiting the benefits of function approximation without the need to construct probability distributions, a task that is recognized as complex and unstable, especially in high dimensions.

3 Intersectional Generalized Disparate Impact

In this section, we recall the original GeDI definition from our previous work [14], and build on that by slightly modifying its semantics to handle multivariate inputs. Differently from any other alternative subgroup fairness definition available in the literature, our estimation procedure inspired by [16] employs continuous function approximation rather than frequency counts, allowing both to capture unfairness across all possible subgroups as well as handling a mixture of categorical and numeric protected features.

3.1 Multivariate GeDI Definition

Let $x, y \in \mathbb{R}^n$ denote the (univariate) protected input and the output target vectors, respectively. Given a mapping function $F : \mathbb{R} \mapsto \mathbb{R}^k$, the Generalized Disparate Impact was originally defined as:

$$\text{GeDI}(x, y; F) = \max_{\alpha} \frac{\text{cov}(F(x) \cdot \alpha, y)}{\sigma(F(x) \cdot \alpha)^2} \quad \text{s.t.} \quad \|\alpha\|_1 = 1 \tag{8}$$

where $\alpha \in \mathbb{R}^k$ is the mixing coefficient vector which maximizes the correlation between the matrix $F(x)$ and the target vector y.

Extending Equation (8) to the multivariate case entails defining a mapping function $F : \mathbb{R}^p \mapsto \mathbb{R}^h$ for an input matrix $X_P \in \mathbb{R}^{n \times p}$. Albeit theoretically simple, such definition is practically challenging for two main reasons. First, designing F may not be straightforward; whereas one-hot encoding the combinatorial product of the protected inputs is sufficient in case of categorical variables only, having at least one numeric attribute would require accounting for joint non-linear effects. Second, the resulting matrix $F(X_P) \in \mathbb{R}^{n \times h}$ would have a higher dimensionality $h \gg k$ with respect to the univariate case, with negative effects on computational efficiency and increased risk of overfitting due to a larger solution space.

Inspired by [16], we rely instead on a neural network to map the protected matrix X_P into a latent space optimizing correlation with the target y. Therefore, we obtain:

$$\begin{aligned} \text{GeDI}(X_P, y; f) = \max_{\theta} \frac{\text{cov}(f(X_P; \theta), y)}{\sigma(f(X_P; \theta))^2} \quad \text{s.t.} \\ \sigma(f(X_P; \theta)) = 1 \end{aligned} \tag{9}$$

where f is a neural network with learnable parameters θ.

Note that we had to revise the original constraint for compatibility with our new estimation technique. Such constraint was designed to limit the sparsity of the mapping function and prevent unbounded results, while also maintaining the link between GeDI and the Disparate Impact Discrimination Index (DIDI) introduced in [1]. Our updated version retains these goals. Indeed, likewise [14], our indicator matches the exact value $\left|\text{cov}(x, y)/\sigma(x)^2\right|$ of the DIDI when fed with a binary input $X_P = x \in \{0, 1\}^n$ and continuous target $y \in \mathbb{R}^n$; moreover, it also guarantees bounded solutions by fixing the standard deviation of the latent vector. We underline that such constraint is model-agnostic, i.e., consistent across different estimation techniques, included the one from [14]. Furthermore, it has a much more understandable semantics, as it directly constrains the sparsity of the mapped vector $f(X_P)$ rather than a custom parameter like α.

Given a matrix $X = \begin{bmatrix} X_P & X_Q \end{bmatrix}$ with protected features P and predictive features Q, we say that a dataset $D = (X, y)$ is γ-fair if:

$$\text{GeDI}(X_P, y; f) \leq \gamma \tag{10}$$

and, similarly, a predictive model $\mathcal{M}$ is γ-fair if the corresponding dataset $D_{\mathcal{M}} = (X, \mathcal{M}(X))$ is γ-fair.

Algorithm 1 Multivariate GeDI Estimation Procedure

Require: $X_P \in \mathbb{R}^{n \times p}$, $y \in \mathbb{R}^n$, $E \in \mathbb{N}$, $\epsilon_i \in \mathbb{R}$, f
  $\theta \leftarrow$ init()
  for $i \in \{1, \ldots, E\}$ **do**
    $\hat{y}(\theta) \leftarrow f(X_P; \theta)$
    $\mathcal{L}(\theta) \leftarrow \frac{\text{cov}(\hat{y}(\theta), y)}{\sigma(\hat{y}(\theta))}$
    $\theta \leftarrow \theta + \epsilon_i \cdot \frac{\partial \mathcal{L}(\theta)}{\partial \theta}$
  end for
  $\hat{y}(\theta) \leftarrow f(X_P; \theta)$
  return $\frac{\text{cov}(\hat{y}(\theta), y)}{\sigma(\hat{y}(\theta))}$

3.2 Multivariate GeDI Computation

On a first sight, it may seem that solving Equation (9) needs a constrained learning procedure. However, simple algebraic manipulation allows to obtain the constraint on the standard deviation for free.

Suppose to have the optimal solution θ^* for which $f(X_P; \theta^*)$ already satisfies the constraint. Then, there exist an infinite number of functions $f(X_P; \theta)$ such that:

$$f(X_P; \theta^*) = \frac{f(X_P; \theta) - \mu(f(X_P; \theta))}{\sigma(f(X_P; \theta))} \tag{11}$$

for which $\sigma(f(X_P; \theta^*)) = 1$ is implied by construction.

By substituting $f(X_P; \theta^*)$ into Equation (9), we get that:

$$\begin{aligned} \text{GeDI}(X_P, y; f) = \frac{\text{cov}(f(X_P; \theta^*), y)}{\sigma(f(X_P; \theta^*))^2} = \\ = \frac{\text{cov}(\frac{f(X_P;\theta) - \mu(f(X_P;\theta))}{\sigma(f(X_P;\theta))}, y)}{\sigma(\frac{f(X_P;\theta) - \mu(f(X_P;\theta))}{\sigma(f(X_P;\theta))})^2} \end{aligned} \tag{12}$$

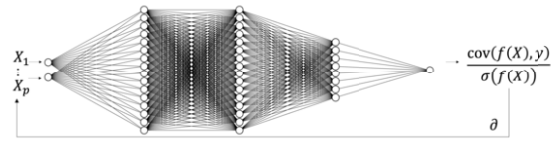
and, since the variance and covariance operators are invariant to the mean, we can discard these terms and move the standard deviation terms outside, further simplifying it to:

$$\begin{aligned} \text{GeDI}(X_P, y; f) = \frac{\sigma(f(X_P; \theta))^{-1} \cdot \text{cov}(f(X_P; \theta), y)}{\sigma(f(X_P; \theta))^{-2} \cdot \sigma(f(X_P; \theta))^2} = \\ = \frac{\text{cov}(f(X_P; \theta), y)}{\sigma(f(X_P; \theta))} \end{aligned} \tag{13}$$

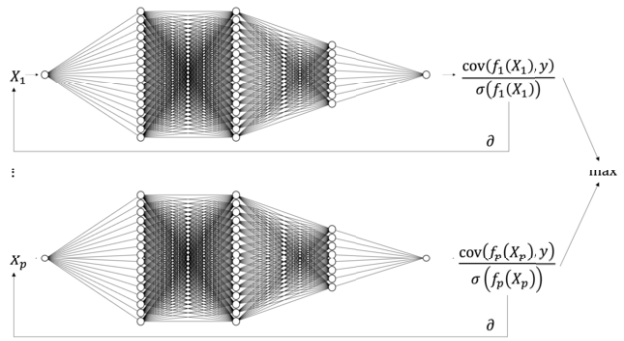
The overall computational procedure is outlined in Algorithm 1. We use this last formulation of GeDI to build a custom loss $\mathcal{L}$ for training our neural network f via gradient ascent steps over E epochs with (variable) learning rate ϵ_i. Once the optimal parameters are retrieved, we map the matrix X_P into the projected vector $f(X_P; \theta)$ and use it to estimate GeDI. Given that the mapping function f performs differentiable operations only, we are also guaranteed that the result retains gradient information, and can therefore be used as a loss regularizer in predictive tasks, as demonstrated in Section 4.

3.3 Properties on Intersectional Subgroup Fairness

In [35], Yang et al. outline four levels of statistical group fairness, of which a graphical example is provided in Figure 1. The simplest, or **Unrestricted Case**, represents the main interest of algorithmic fairness papers as it only considers a single protected attribute. Following, the **Independent Groups** level takes into account multiple protected attributes, although treating them independently and therefore ignoring combined discrimination effects. In the **Intersectional Groups** level, combined effects on joint attributes are also examined, despite overlooking possible discrimination on a subgroup



(a) Intersectional GeDI Architecture.



(b) Independent GeDI Architecture.

Figure 2. Neural architectures for GeDI mapping functions.

granularity by neglecting marginal probabilities during the computation. Finally, the most comprehensive scenario is the **Gerrymandering Groups** level, named after the term "Fairness Gerrymandering" coined in [20], where the authors build an intersectionally-aware classifier that ensures fair policies across all protected subgroups.

The original definition of Statistical Parity Subgroup Fairness (SPSF) proposed in [20] consists in enumerating all possible (non-empty) subgroups to assess unfairness by comparing subgroup acceptance rates to the dataset average. Although exhaustive, this leads in practice to a super-exponential number of elements. Foulds et al. [12] address this issue by proposing a different indicator, Empirical Differential Fairness (EDF), which ensures that if a larger set of protected values meets a statistical fairness threshold ϵ, all its subsets do too. In principle, EDF reduces the number of groups to consider, counting to an exponential rather than super-exponential number; yet, it keeps facing the same challenge of SPSF, namely the potential under/over-constrainment of underrepresented subgroups.

We solve both issues using a latent space projection, which helps avoiding enumeration and achieving smoother results. By using a neural network, we can balance the bias-variance trade-off, hence reducing overfitting and providing some information for underrepresented groups through continuous approximation. Similarly to EDF, our multivariate GeDI indicator encompasses all marginal subgroups; i.e., if a dataset $D = (\begin{bmatrix} X_P & X_Q \end{bmatrix}, y)$ is γ-fair, then any dataset $D' = (\begin{bmatrix} X'_P & X_Q \end{bmatrix}, y)$ with protected matrix X'_P containing a subset of the whole protected features of X_P will be γ-fair as well. In fact, we demonstrate that our indicator can be restricted to any of the three (sub-)levels from [35] by constraining the mapping function f to a specific subspace $\mathcal{H} \subseteq \mathbb{R}^p \mapsto \mathbb{R}$. For the **Unrestricted Case**, this involves selecting a univariate function limited to the only protected attribute X_i considered, i.e.:

$$\mathcal{H}_{\text{unr}} = \{f \mid \exists g \in \mathbb{R} \mapsto \mathbb{R} \quad \text{s.t.} \quad f(X_P) = g(X_i)\} \tag{14}$$

Similarly, fairness on **Independent Groups** can be assessed by aggregating the outcome of p distinct univariate functions, each focusing on a single protected attribute only, e.g.:

$$\mathcal{H}_{\text{ind}} = \{f \mid \exists g \in \mathbb{R} \mapsto \mathbb{R} \quad \text{s.t.} \quad f(X_P) = \max\{g(X_i)\}\} \tag{15}$$

Algorithm 2 Intersectionally-Aware Predictive Model Training

Require: $X = \begin{bmatrix} X_P \in \mathbb{R}^{n \times p} & X_Q \in \mathbb{R}^{n \times q} \end{bmatrix}, \quad y \in \mathbb{R}^n$
Require: $E \in \mathbb{N}, \quad \epsilon_i^{\mathcal{N}}, \epsilon_i^{\lambda} \in \mathbb{R}, \quad \tau \in \mathbb{R}, \quad \mathcal{I}, \quad \mathcal{L}_T, \quad \mathcal{N}$
  $\lambda \leftarrow 0$
  $\theta \leftarrow$ init()
  for $i \in \{1, \dots, E\}$ **do**
    ▷ *Update θ (Gradient Descent)*
    $\hat{y}(\theta) \leftarrow \mathcal{N}(X; \theta)$
    $\mathcal{L}(\lambda, \theta) \leftarrow \mathcal{L}_T(\hat{y}(\theta), y) + \lambda \cdot \max\{\mathcal{I}(X_P, \hat{y}(\theta)) - \tau, 0.0\}$
    $\theta \leftarrow \theta - \epsilon_i^{\mathcal{N}} \cdot \frac{\partial \mathcal{L}(\lambda, \theta)}{\partial \theta}$
    ▷ *Update λ (Gradient Ascent)*
    $\hat{y}(\theta) \leftarrow \mathcal{N}(X; \theta)$
    $\mathcal{L}(\lambda, \theta) \leftarrow \mathcal{L}_T(\hat{y}(\theta), y) + \lambda \cdot \max\{\mathcal{I}(X_P, \hat{y}(\theta)) - \tau, 0.0\}$
    $\lambda \leftarrow \lambda + \epsilon_i^{\lambda} \cdot \frac{\partial \mathcal{L}(\lambda, \theta)}{\partial \lambda}$
  end for
  return θ

Proving the same for the **Intersectional Groups** level is more complex. We restrict our analysis to categorical protected attributes only, underlining that this is consistent with all the existing literature. Suppose that we have p protected attributes, each taking a value $v_i \in \{1, \dots, d_i\}$; we define $s_i \in \{0, 1\}^{d_i}$ as the *selector* vector consisting in the one-hot encoding of each attribute v_i. We can now optimize a p-dimensional tensor $F \in \mathbb{R}^{d_1 \times \cdots \times d_p}$ capturing an output for each combination of protected values, thus:

$$\mathcal{H}_{\text{int}} = \left\{ f \mid \exists F \text{ s.t. } f(X_P) = \left(\left(F^T \cdot s_1 \right)^T \cdot \dots \right)^T \cdot s_p \right\} \tag{16}$$

where the selector vectors precisely select the output linked to the protected values v_i. Given that we do not impose any specific constraint to our function except for the model bias of the neural architecture itself, it follows that our indicator implicitly ties to the **Gerrymandering Groups** level, and that achieving γ-fairness automatically ensures γ-fairness on each level.

4 Empirical Evaluation

We conduct two experiments on four established benchmarks for algorithmic fairness in order to empirically validate our claims. Specifically, in Section 4.1 we compare our multivariate GeDI indicator (Figure 2(a)) against multiple univariate indicators (Figure 2(b)) to demonstrate its ability to capture intersectional effects. Thereafter, in Section 4.2 we compare GeDI with two other indicators for intersectional statistical parity available in the literature, i.e., Subgroup Fairness Statistical Parity (SFSP) and Empirical Differential Fairness (EDF), to prove its stronger expressiveness and robustness.

We implement our code using Python 3.12, running all the experiments on a MacBook Pro with an Apple M3 Pro processor and 36GB RAM, without resorting to GPUs. To ensure reproducibility and deterministic neural network behavior, we initialize seeds for random operations using Pytorch Lightning's `seed_everything` method and the `deterministic=True` training option prior to execution. The code to replicate the experiments is publicly available at the following link: ***https://github.com/giuluck/intersectional-fairness***.

We outline our constrained training procedure in Algorithm 2. We leverage a feed-forward neural network $\mathcal{N}$ with two 32-unit hidden layers using ReLU activation and train it full-batch for 500 epochs. The output unit has either sigmoid or linear activation depending

Regularizer	Task Loss ($\times 10^2$) train	Task Loss ($\times 10^2$) val	% Inter train	% Inter val	% Indep train	% Indep val	Time (s)
Dataset: Student (649 × 43)	**Protected**: sex ∈ { *F, M* } school ∈ { *GP, MS* } address ∈ { *Urban, Rural* } age ∈ $\mathbb{R}$						
//	**0.42 ± 0.11**	2.92 ± 0.77	89 ± 11	104 ± 04	91 ± 18	82 ± 24	**02 ± 00**
Inter	1.64 ± 0.16	**2.60 ± 0.84**	**11 ± 02**	**65 ± 07**	**12 ± 03**	**27 ± 08**	21 ± 00
Indep	0.75 ± 0.09	2.91 ± 0.90	58 ± 13	90 ± 14	27 ± 25	52 ± 22	74 ± 00
Dataset: Compas (6172 × 12)	**Protected**: Gender ∈ { *F, M* } Race ∈ { *African_American, Asian, Hispanic, Native_American, White, Other* }						
//	**59.34 ± 0.28**	**61.10 ± 1.21**	99 ± 02	100 ± 11	99 ± 05	100 ± 13	**15 ± 00**
Inter	61.09 ± 0.30	61.96 ± 1.08	**19 ± 01**	**23 ± 11**	**19 ± 02**	**21 ± 10**	59 ± 00
Indep	60.92 ± 0.31	61.83 ± 1.06	32 ± 02	34 ± 08	20 ± 01	23 ± 06	322 ± 01
Dataset: Law (20512 × 12)	**Protected**: gender ∈ { *F, M* } race ∈ { *asian, black, hisp, white, other* } fam_inc ∈ $\mathbb{R}$						
//	**14.37 ± 0.17**	**14.81 ± 0.52**	97 ± 02	100 ± 05	101 ± 03	102 ± 06	**55 ± 02**
Inter	15.67 ± 0.08	15.89 ± 0.37	**19 ± 00**	**22 ± 03**	**19 ± 00**	**19 ± 03**	143 ± 02
Indep	15.68 ± 0.08	15.89 ± 0.40	33 ± 02	35 ± 04	20 ± 01	22 ± 03	647 ± 04
Dataset: Adult (45222 × 103)	**Protected**: gender ∈ { *F, M* } race ∈ { *Amer-Indian-Eskimo, Asian-Pac-Islander, Black, White, Other* } age ∈ $\mathbb{R}$						
//	**9.18 ± 0.04**	**10.37 ± 0.22**	101 ± 01	102 ± 02	144 ± 02	145 ± 06	**141 ± 01**
Inter	11.78 ± 0.09	12.23 ± 0.33	**19 ± 02**	**25 ± 02**	24 ± 06	30 ± 11	437 ± 09
Indep	11.74 ± 0.07	12.05 ± 0.31	34 ± 01	38 ± 01	**20 ± 01**	**28 ± 08**	1694 ± 67

Table 1. Comparison between the Intersectional and Independent definitions of GeDI on four benchmark datasets. We report mean and standard deviation from a 5-fold cross-validation procedure across three models: unconstrained (//), and constrained using either the Intersectional or the Independent indicator up to 20% of its value in the original data. For each dataset, we also show its shape and protected features domains. Best values are highlighted in bold.

on whether it is adopted for a classification or a regression task, respectively; accordingly, we use binary cross-entropy (BCE) or mean squared error (MSE) as task loss $\mathcal{L}_T$. We adopt Adam as network optimizer, with initial learning rate $\epsilon_1^{\mathcal{N}} = 0.001$, and enforce fairness constraints via a lagrangian multiplier λ balancing the regularization term, leading to the following training loss:

$$\mathcal{L}_T(y, \mathcal{N}(X;\theta)) + \lambda \cdot \max\{\mathcal{I}(X_P, \mathcal{N}(X;\theta)) - \tau, 0.0\} \quad (17)$$

where $\mathcal{I}$ is the fairness indicator, X_P the protected feature matrix, and τ the constraint threshold. Rather than relying on a fixed multiplier, we automatically adjust its value during training via alternate gradient ascent steps as described in [11]. A separate optimizer, configured in the same way as the network one, is used to handle the learning rates ϵ_i^{λ}; this eliminates the need of a manual tuning phase, hence yielding more accurate and less computationally expensive results. Moreover, as noted in [20], practical use cases involve stakeholders or policymakers to which the choice of specific thresholds is delegated, hence this method allows to satisfy the requirements while avoiding both under- and over-constraining.

Regarding the benchmarks, we chose the four tabular datasets mentioned in [15], namely: COMPAS[2], Student Performance[3], Law School Admission[4], and Adult Income[5]. Except for *Student*, which has a continuous target, all the other benchmarks involve a binary classification task. Independently from the task, we preprocess each dataset by removing duplicate or non-predictive features, normalizing the target within $[0, 1]$, standardizing continuous input features, and one-hot encoding multi-class categorical attributes.

4.1 Intersectional vs. Independent GeDI

We start by examining the gap between fairness as defined by our multivariate (Intersectional) GeDI and a univariate approach that estimates it on each protected attribute independently (Independent). Recalling the four levels described in Section 3.3, these indicators correspond to the **Gerrymandering Groups** and the **Independent Groups** levels, respectively.

GeDI is computed according to Algorithm 1. We adopt the same neural architecture from [16], i.e., a feed-forward network with three hidden layers (16, 16, and 8 units). Layers are initialized with Xavier uniform weights and a constant bias term (0.01), and are followed by a ReLU activation. For the Intersectional indicator, we build a single network with input dimension p[6]; as regards the Independent indicator, instead, p different networks with univariate input are used, and the maximum estimated GeDI is eventually taken as result – see Figure 2. In both cases, neural networks are trained full-batch for 1000 steps using an Adam optimizer with learning rate 0.0005. As per [16], when the indicator is used as regularizer within a learning procedure, the previously computed mapping network f_i is used as a warm start for training epoch $i + 1$ and fine-tuned for 50 steps only.

Table 1 presents the results obtained on four benchmarks from an unconstrained model (//) and two models constrained up to a threshold τ_i computed as the 20% of the estimation on the original dataset – namely $\tau_i = 0.2 \cdot \mathrm{GeDI}_i(X_P, y; f)$, for $i \in \{\mathrm{Inter}, \mathrm{Indep}\}$. We use percentages for better comparison across indicators, underlining that this has no impact on our evaluation since the original dataset estimate is fixed, hence results would just be scaled by a constant factor. We run a 5-fold cross-validation procedure, reporting mean and standard deviation on both training and validation sets for: (*i*) the task loss computed on the final predictions, (*ii*) each considered indicator, and (*iii*) computational times. Best results are highlighted in bold. As task loss, we use mean-squared error for the *Student* dataset and binary cross-entropy for all the others. We also report the dataset shape ($n \times m$) and its set of protected features along with their domains, which can be binary, multi-class, or continuous.

Results align with our theoretical claims. Both the Intersectional and the Independent indicators mostly meet the 20% threshold on their respective metric in the training split, confirming a correct constraint enforcement. Albeit no guarantees can be provided on valida-

[2] www.kaggle.com/datasets/danofer/compass
[3] www.kaggle.com/datasets/larsen0966/student-performance-data-set
[4] www.kaggle.com/datasets/danofer/law-school-admissions-bar-passage
[5] www.kaggle.com/datasets/wenruliu/adult-income-dataset

[6] Note that multi-class categorical features in $\{1, \ldots, d_i\}$ are one-hot encoded, thus contributing as d_i different features.

Regularizer	Task Loss ($\times 10^2$)		% GeDI		% EDF		% SPSF		Time (s)
	train	val	train	val	train	val	train	val	
Dataset: Compas (6172 × 12)	**Protected**:	Gender ∈ { *F, M* }	Race ∈ { *African_American, Asian, Hispanic, Native_American, White, Other* }						
//	**59.34 ± 0.28**	**61.10 ± 1.21**	99 ± 02	100 ± 11	91 ± 17	102 ± 27	99 ± 04	100 ± 11	**15 ± 00**
GeDI	61.09 ± 0.30	61.96 ± 1.08	**19 ± 01**	**23 ± 11**	25 ± 03	**34 ± 07**	**19 ± 01**	**20 ± 12**	59 ± 00
EDF	59.93 ± 0.33	60.86 ± 1.27	77 ± 03	79 ± 11	**20 ± 00**	50 ± 13	81 ± 04	82 ± 11	22 ± 00
SPSF	60.81 ± 0.28	61.94 ± 0.98	33 ± 04	37 ± 08	68 ± 11	70 ± 13	**19 ± 01**	22 ± 09	22 ± 00
Dataset: Law (20512 × 12)	**Protected**:	gender ∈ { *F, M* }	race ∈ { *asian, black, hisp, white, other* }						
//	**14.37 ± 0.17**	**14.81 ± 0.52**	101 ± 02	102 ± 05	102 ± 04	104 ± 09	102 ± 04	102 ± 04	**55 ± 02**
GeDI	15.83 ± 0.09	16.06 ± 0.30	**19 ± 01**	**20 ± 03**	**18 ± 01**	**19 ± 02**	**20 ± 01**	**20 ± 03**	142 ± 01
EDF	15.44 ± 0.08	15.66 ± 0.43	29 ± 00	29 ± 04	20 ± 00	25 ± 05	32 ± 01	32 ± 04	76 ± 00
SPSF	15.51 ± 0.08	15.72 ± 0.40	30 ± 01	31 ± 04	37 ± 02	38 ± 07	**20 ± 01**	22 ± 03	77 ± 00
Dataset: Adult (45222 × 103)	**Protected**:	gender ∈ { *F, M* }	race ∈ { *Amer-Indian-Eskimo, Asian-Pac-Islander, Black, White, Other* }						
//	**9.18 ± 0.04**	**10.37 ± 0.22**	100 ± 02	100 ± 04	100 ± 03	116 ± 26	100 ± 02	100 ± 04	**141 ± 01**
GeDI	10.34 ± 0.06	10.91 ± 0.27	**17 ± 03**	**19 ± 04**	**12 ± 02**	**26 ± 09**	18 ± 02	18 ± 06	430 ± 06
EDF	10.05 ± 0.08	10.56 ± 0.25	40 ± 01	41 ± 03	20 ± 01	34 ± 08	42 ± 01	42 ± 03	187 ± 00
SPSF	10.39 ± 0.03	10.87 ± 0.27	31 ± 01	31 ± 04	42 ± 01	48 ± 07	**16 ± 02**	**17 ± 04**	188 ± 00

Table 2. Comparison between GeDI and two other intersectional statistical parity indicators, EDF and SPSF. We report mean and standard deviation from a 5-fold cross-validation procedure on three benchmark datasets across four models: unconstrained (//), and constrained using one of the tested indicators up to 20% of its value in the original data. For each dataset, we also show its shape and protected feature domains. Best values are highlighted in bold.

tion splits as for any other statistical fairness constraint, we see that validation unfairness only slightly increases in all benchmarks except for *Student*, likely due to its low cardinality. Most importantly, the Intersectional constraint consistently yields better outcomes with respect to the Independent one across *all* fairness metrics, demonstrating its ability to capture more forms of discrimination at the expenses of little to no increment in the task loss. This suggests that enforcing intersectional fairness would not impact the fairness-accuracy trade-off more than having independent constraints. Finally, the Intersectional indicator is much faster to compute respectively to its Independent alternative, as the latter involves training multiple neural networks rather than just one.

4.2 Baselines Comparison

After proving that our multivariate GeDI effectively captures intersectional discrimination, we compare it with two other intersectional statistical parity indicators, namely Empirical Differential Fairness (EDF) and Statistical Parity Subgroup Fairness (SPSF).

We follow the experimental setup of Section 4.1, running a 5-fold cross-validation procedure for both the unconstrained (//) and the three constrained models, albeit disregarding both the *Student* benchmark and the continuous protected attributes from other benchmarks since EDF and SPSF can only handle categorical protected inputs and binary targets. In order to have differentiable constraints, we implement EDF and SPSF using `torch` primitives according to their definition in [12] and [20], respectively:

$$\mathrm{EDF}(X, y) = \max_S \left\{ \log\left(\frac{N_S^+}{N_S}\right) \right\} - \min_S \left\{ \log\left(\frac{N_S^+}{N_S}\right) \right\} \quad (18)$$

$$\mathrm{SPSF}(X, y) = \max_S \left\{ \frac{N_S}{N} \cdot \left| \frac{N_S^+}{N_S} - \frac{N^+}{N} \right| \right\} \quad (19)$$

where $S \in \mathcal{S}$ is any non-empty subgroup of protected attributes; N and N^+ represent the total number of records in the dataset, and the records with a positive outcome $y_i = 1$, respectively; and N_S and N_S^+ indicate the respective quantities conditioned on subgroup S.

Results are shown in Table 2. As for the previous experiment, each indicator satisfies the 20% threshold on its respective training split measure, confirming the validity of our learning model choice. Additional outcomes meet our expectations, with the *GeDI constraint performing well even when measured against other indicators* – especially in the Law benchmark, where it consistently yields optimal values across each metric; conversely, EDF and SPSF constrains perform poorly when evaluated both on GeDI and on their counterpart, proving that *GeDI has a greater ability to capture discriminative effects*. Another hint at GeDI's expressivity is its stronger robustness in validation results, showing that its training-time enforcement is better mirrored at test time with respect to alternative indicators, at the expenses of a negligible increment in task loss. Nonetheless, due to its reliance on neural approximation, GeDI requires a longer computational time as opposed to any other enumeration-based method.

5 Conclusions

We introduced a multivariate definition of our Generalized Disparate Impact (GeDI) indicator introduced in [14], along with a computational estimation algorithm inspired by [16] which employs a neural network to optimize correlation between a univariate latent space projection of the protected input matrix and the target vector.

We demonstrated that, when used as a fairness measure, such multivariate GeDI definition is able to intercept unfairness at both the independent and the intersectional levels, inherently accounting for any kind of subgroup fairness gerrymandering. Unlike other measures available in the literature, GeDI relies on approximation techniques rather than enumeration, allowing it to both handle continuous protected attributes and better manage underrepresented subgroups.

Empirical evaluations conducted on four benchmark datasets confirm our theoretical claims. Multivariate GeDI is in fact an effective intersectional fairness measure, outperforming alternative indicators on fairness enforcement tasks despite a minimal loss in predictive accuracy. We reckon that enforcing algorithmic fairness requires analyzing social oppression and power dynamics, and that any research on the topic is insufficient unless tied to real-world actions as mentioned in [23] and [29]; nonetheless, disposing of technically comprehensive indicators is a necessary step to reach this goal, and we hope that our work could help advancing towards it.

Ethics Statement

Despite the ethical nature of our subject, we declare no ethical concerns in our work due to its theoretical focus. Moreover, we remark that no claims on how automated intersectional discrimination impacts individuals and society can be drawn from this contribution, as its analysis is restricted to fairness measurement and enforcement. Notwithstanding the absence of explicit ethical implications, we recall that the term "intersectionality" cannot be untied from its systemic roots generating oppression and discrimination. Therefore, we hope that this paper could stimulate more profound consideration on the ethical aspects of data-driven predictive and prescriptive systems.

References

[1] S. Aghaei, M. J. Azizi, and P. Vayanos. Learning optimal and fair decision trees for non-discriminative decision-making. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, volume 33, pages 1418–1426. AAAI Press, July 2019.

[2] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias. In *Ethics of data and analytics*, pages 254–264. Auerbach Publications, 2022.

[3] D. Atewologun. Intersectionality theory and practice. *Oxford Research Encyclopedia of Business and Management*, 2018.

[4] R. Binns. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 514–524. Association for Computing Machinery, 2020. doi: 10.1145/3351095.3372864.

[5] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In S. A. Friedler and C. Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR, 23–24 Feb 2018.

[6] K. Crenshaw. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. In *Feminist legal theories*, pages 23–51. Routledge, 2013.

[7] F. Ding, M. Hardt, J. Miller, and L. Schmidt. Retiring adult: new datasets for fair machine learning. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21. Curran Associates Inc., 2021.

[8] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. S. Zemel. Fairness through awareness. In S. Goldwasser, editor, *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*, pages 214–226. ACM, 2012.

[9] A. Fabris, N. Baranowska, M. J. Dennis, D. Graus, P. Hacker, J. Saldivar, F. Zuiderveen Borgesius, and A. J. Biega. Fairness and bias in algorithmic hiring: a multidisciplinary survey. *ACM Transactions on Intelligent Systems and Technology*, 2023.

[10] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In L. Cao, C. Zhang, T. Joachims, G. I. Webb, D. D. Margineantu, and G. Williams, editors, *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*, pages 259–268. ACM, 2015.

[11] F. Fioretto, P. V. Hentenryck, T. W. K. Mak, C. Tran, F. Baldo, and M. Lombardi. Lagrangian duality for constrained deep learning. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track*, pages 118–135. Springer International Publishing, 2021.

[12] J. R. Foulds, R. Islam, K. N. Keya, and S. Pan. An intersectional definition of fairness. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 1918–1921, 2020. doi: 10.1109/ICDE48307.2020.00203.

[13] A. Ghosh, L. Genuit, and M. Reagan. Characterizing intersectional group fairness with worst-case comparisons. In *AIDBEI*, 2021.

[14] L. Giuliani, E. Misino, and M. Lombardi. Generalized disparate impact for configurable fairness solutions in ML. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 11443–11458. PMLR, July 2023.

[15] U. Gohar and L. Cheng. A survey on intersectional fairness in machine learning: Notions, mitigation, and challenges. 05 2023. doi: 10.24963/ijcai.2023/742.

[16] V. Grari, S. Lamprier, and M. Detyniecki. Fairness-aware neural rényi minimization for continuous features. In C. Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 2262–2268. International Joint Conferences on Artificial Intelligence Organization, July 2020.

[17] U. Hebert-Johnson, M. Kim, O. Reingold, and G. Rothblum. Multicalibration: Calibration for the (Computationally-identifiable) masses. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1939–1948. PMLR, 10–15 Jul 2018.

[18] High-Level Expert Group on AI. Ethics guidelines for trustworthy AI. Technical report, European Commission, Apr. 2019.

[19] Z. Jiang, X. Han, C. Fan, F. Yang, A. Mostafavi, and X. Hu. Generalized demographic parity for group fairness. In *International Conference on Learning Representations*, 2022.

[20] M. Kearns, S. Neel, A. Roth, and Z. S. Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2564–2572. PMLR, 10–15 Jul 2018.

[21] R. F. Kizilcec and H. Lee. Algorithmic fairness in education. In *The ethics of artificial intelligence in education*, pages 174–202. Routledge, 2022.

[22] I. Kong, K. Kim, and Y. Kim. Fair representation learning for continuous sensitive attributes using expectation of integral probability metrics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3784–3795, 2025. doi: 10.1109/TPAMI.2025.3538915.

[23] Y. Kong. Are "intersectionally fair" ai algorithms really fair to women of color? a philosophical analysis. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 485–494. ACM, June 2022. doi: 10.1145/3531146.3533114.

[24] D. H. Locke, B. Hall, J. M. Grove, S. T. Pickett, L. A. Ogden, C. Aoki, C. G. Boone, and J. P. O'Neil-Dunne. Residential housing segregation and urban tree canopy in 37 us cities. *NPJ Urban Sustainability*, 1(1): 15, 2021.

[25] T. Madiega. Artificial intelligence act. *European Parliament: European Parliamentary Research Service*, 2021.

[26] J. Mary, C. Calauzènes, and N. E. Karoui. Fairness-aware learning for continuous attributes and treatments. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4382–4391. PMLR, June 2019.

[27] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys*, 54(6):1–35, July 2021.

[28] M. Molina and P. Loiseau. Bounding and approximating intersectional fairness through marginal fairness. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22. Curran Associates Inc., 2022.

[29] J. Morley, L. Kinsey, A. Elhalal, F. Garcia, M. Ziosi, and L. Floridi. Operationalising ai ethics: barriers, enablers and next steps. *AI Soc.*, 38 (1):411–423, Nov. 2021.

[30] A. Rényi. On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungarica*, 10:441–451, 1959.

[31] G. Rothblum and G. Yona. Probably approximately metric-fair learning. In *International Conference on Machine Learning*, 2018.

[32] C. Simoiu, S. Corbett-Davies, and S. Goel. The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics*, 11:1193–1216, 09 2017. doi: 10.1214/17-AOAS1058.

[33] D. Ueda, T. Kakinuma, S. Fujita, K. Kamagata, Y. Fushimi, R. Ito, Y. Matsui, T. Nozaki, T. Nakaura, N. Fujima, et al. Fairness of artificial intelligence in healthcare: review and recommendations. *Japanese Journal of Radiology*, 42(1):3–15, 2024.

[34] A. Wang, V. V. Ramaswamy, and O. Russakovsky. Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 336–349. ACM, June 2022. doi: 10.1145/3531146.3533101.

[35] F. Yang, M. Cisse, and S. Koyejo. Fairness with overlapping groups. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20. Curran Associates Inc., 2020.