A Generalised Exponentiated Gradient Approach to Enhance Fairness in Binary and Multi-class Classification Tasks

Maryam Boubekraoui[a], Giordano d'Aloisio[b], Antinisca Di Marco[b]

[a] *Department One, Address One, City One, 00000, State One, Country One*
[b] *Department of Information Engineering, Computer Science and Mathematics, University of L'Aquila, L'Aquila, Italy*

Abstract

The wide adoption of AI- and ML-based systems in sensitive domains raises severe concerns about their fairness. The research community has proposed several methods for bias mitigation in recent years. However, despite its relevance and wide application, the topic of bias mitigation in multi-class classification settings – i.e., where the number of classes to predict is > 2 – remains under-explored.

To address this limitation, in this paper, we tackle the problem of fairness in multi-class classification settings. We first formulate the problem of fair learning in multi-class classification as a multi-objective problem between effectiveness (i.e., prediction correctness) and multiple linear fairness constraints. Next, we propose a Generalised Exponentiated Gradient (GEG) algorithm to solve this task. GEG is an in-processing algorithm that enhances fairness in binary and multi-class classification settings under multiple fairness definitions. We conduct an extensive empirical evaluation of GEG and demonstrate that it is a practical solution for bias mitigation without compromising prediction effectiveness. Additionally, from our empirical evaluation, we draw a set of practical insights for adopting GEG in real-world scenarios. GEG is general and flexible, making it applicable to multiple use-case scenarios.

Keywords: bias, fairness, multi-class classification, classification task

---

1. Introduction

With the increasing adoption of AI- and ML-based software systems in sensitive domains such as healthcare [1], finance [2], and education [3], it is critical to ensure that they act in an *unbiased* and *ethical* way. In other words, they must be *fair*. The relevance of *software fairness* has been highlighted in recent years not only in research literature [4, 5, 6], but also in regulations such as the European Union's recently introduced AI Act [7].

Bias can be defined as the systematic discrimination or favouritism of a software system toward individuals or groups identified by a set of sensitive features [4]. When not properly addressed, it may cause severe consequences and discrimination, like for the recruitment instrument employed by Amazon, which penalised women candidates for IT job positions,[1] or the criminal recidivism predictions made by the commercial risk assessment software COMPAS, which misjudged black individuals based on biased profiling [8].

For this reason, the research community has proposed several methods for *bias mitigation* at different processing levels [4, 5]. However, the majority of them can be applied only to binary classification tasks. Instead, several examples of multi-class classification approaches have been applied in sensitive domains such as education [9, 10], food [11, 12], and health [13]. Ensuring that these systems behave in a *fair* and unbiased way is paramount, also to achieve some of the United Nations Sustainable Development Goals (SDG) [14], e.g., SDG 2 (zero huger) [11, 12], SDG 3 (good health and well-being) [13], and SDG 4 (quality education) [9, 10].

To address this limitation, in this paper, we first formulate the problem of fairness in multi-class classification as a multi-objective problem between effectiveness (i.e., prediction correctness) and multiple fairness definitions. Next, we propose a Generalised Exponentiated Gradient (GEG) algorithm to solve this task. GEG is an extension of the original Exponentiated Gradient (EG) bias mitigation algorithm first proposed by Agarwal et al. for binary classification [15]. However, GEG differs from the original EG approach in two aspects: first, it can mitigate bias in both binary and multi-class classification tasks; second, it mitigates bias across multiple fairness constraints simultaneously. These additions make GEG more flexible and practical for multiple use cases.

We perform an extensive evaluation of GEG, benchmarking it against

[1] `https://www.bbc.co.uk/news/technology-45809919`

six approaches across seven multi-class and three binary datasets, using four widely adopted effectiveness metrics and three fairness definitions. Results show that GEG is a practical approach for bias mitigation in multi-class classification, overcoming existing baselines. Additionally, from our empirical evaluation, we draw practical tips on employing GEG in real-case scenarios.

Specifically, the main contributions of our work are the following:

- We formulate the problem of fairness in multi-class classification as a multi-objective problem between multiple fairness constraints.
- We propose a Generalised Exponentiated Gradient (GEG) approach to mitigate bias in binary and multi-class classification tasks under multiple fairness constraints simultaneously.
- We perform an extensive empirical evaluation of GEG against multiple baselines, datasets, and metrics.
- We draw a set of practical insights on adopting GEG in real-case scenarios.
- We release a replication package including a Python implementation of GEG and the results of our empirical evaluation to foster future research [16].

The rest of this paper is structured as follows: Section 2 provides background knowledge on fairness and discusses related work. Section 3 presents the fairness learning in multi-class classification as a multi-objective optimisation problem and describes the GEG approach. Section 4 describes the empirical evaluation performed, while Section 5 discusses the obtained results and provides practical insights. Finally, Section 6 discusses future work and concludes the study.

2. Background and Related Work

Fairness is defined as: *"The absence of prejudice and favouritism of a software system toward individuals or groups"* [4]. When a system behaves unfairly, it is said to be *biased*. Bias can originate from three main sources [4]: the **data** used to train the AI and ML components, a **biased implementation** of the AI and ML components, and the **people** that interact with those components. In this paper, we focus on mitigating bias in an ML model trained on biased data (i.e., *data* bias). More specifically, we focus on multi-class classification with structured, tabular data.

In general, the fairness of an ML model can be assessed following two main criteria: **individual** and **group** fairness [17]. **Individual** fairness requires that two individuals who are similar to one another receive the same treatment (i.e., an ML model should make identical predictions). Most of the time, two individuals are treated as similar if they only differ in sensitive attributes[2] (e.g., *ethnicity*, *gender*, *age*). **Group** fairness, on the other hand, addresses fairness by treating population groups, defined by protected attributes (like *ethnicity*, *gender*, or *age*), equally. In this work, we focus on group fairness criteria, as they are more common and have been more extensively addressed in previous work [18, 19]. Specifically, many group fairness definitions and corresponding metrics have been proposed in the literature [4, 5]. The general idea behind all group fairness definitions is that, given two groups named **privileged** and **unprivileged** (e.g., *men* and *women*), they must have the same probability of having a given **positive** outcome from the ML model, possibly conditioned on the ground truth label [4]. In Sections 3 and 4, we provide the formal specification of the fairness definitions we address in this paper.

In addition to measuring bias, research has proposed several methods for mitigating bias at different processing levels [18, 4]. Generally, improvement in fairness implies a reduction in the effectiveness of a model's predictions [6, 18, 20], and all bias mitigation methods try to identify the optimal trade-off between fairness and effectiveness. In particular, there are three main categories of bias mitigation methods based on when they are applied in an ML workflow: **pre-processing**, **in-processing**, and **post-processing** [21, 18, 4]. **Pre-processing** bias mitigation methods aim to reduce bias by applying changes to the training data. For instance, one can assign more weight to data instances for a population group that is prone to being misclassified [22, 23]. **In-processing** bias mitigation methods make changes to the design and training process of ML models to achieve fairness. One example is the inclusion of fairness metrics as part of the training loss [24, 15]. Alternatives include the tuning of hyperparameters [25] or the use of ensembles, where each model can consider different popula-

[2]In the rest of this paper, we will use the terms "sensitive attributes", "protected attributes" or "sensitive features" as synonyms.

tion groups [26] or metrics [27]. **Post-processing** bias mitigation methods are applied once an ML model has been successfully trained. This can involve changes to the model's predictions [28] or modifications to the model itself [29].

The majority of bias mitigation methods proposed in the literature focus on binary classification tasks, while very few address multi-class classification [18, 4]. One of the first approaches proposed for multi-class bias mitigation is the *Blackbox* post-processing approach by Putzel et al. [30], which extends the *Equalized Odds* algorithm [28] to the multi-class setting. This algorithm builds a linear optimisation program that optimises the predictions of an already trained classifier to satisfy the *Equalized Odds* fairness definition for multi-class settings. A similar approach is the *Demographic Parity* post-processing approach proposed by Denis et al. [31], where the predictions are instead optimised under the *Demographic Parity* fairness definition. One of the most recent approaches for multi-class bias mitigation is the pre-processing *Debiaser for Multiple Variables (DEMV)* algorithm proposed by d'Aloisio et al. [23]. This algorithm extends the *Sampling* method of Kamiran et al. [22] to the multi-class setting and has been shown to overcome existing bias mitigation methods for multi-class classification.

Our proposed approach differs from the previous ones in that it is an in-processing bias mitigation method. In particular, our work extends the *Exponentiated Gradient* in-processing algorithm proposed by Agarwal et al. [15]. In their work, the authors formulate a multi-objective optimisation problem to train a binary classifier under specific fairness constraints. Next, they present an Exponentiated Gradient (EG) method to solve this optimisation task. Our proposed approach extends the original EG algorithm to the multi-class classification setting and to the simultaneous optimisation of multiple fairness constraints, making it more general and practical for real-world use cases.

3. Methodology

In this section, a general in-processing fairness-enhancing model for classification tasks is presented that can handle binary as well as multi-class data. This model offers the flexibility to include several fairness metrics in a single optimisation problem. Our goal is to make fair and accurate predictions with minimal loss in prediction effectiveness. Our work is inspired by the widely adopted reduction-based framework of Agarwal et al. [32], extending it to accommodate multi-class classification and additional fairness conditions. These modifications make the model more adaptable and suitable for real-world applications.

Consider training data as triplets (X, A, Y), where $X \in \mathcal{X}$ represents the input features, and $A \in \mathcal{A}$ is a protected attribute like *gender* or *race* that could affect fairness, while $Y \in \mathcal{Y}$ is the output label. The set $\mathcal{Y}$ can be binary (i.e., $\{0, 1\}$) or multi-class (like $\{0, 1, \ldots, K\}$). The idea is to learn a classifier $h : \mathcal{X} \to \mathcal{Y}$ from a hypothesis class $\mathcal{H}$ that meets fairness constraints and gives accurate predictions. We assume that the true labels Y and predicted labels $h(X)$ belong to the same space $\mathcal{Y}$, and define $y_p \in \mathcal{Y}$ as the *positive* or *favourable* outcome.

A common technique for guaranteeing fairness during model training is to include fairness as a constraint in the learning objective [32, 33]. The constrained optimisation problem that results from this is as follows:

$$\min_{h \in \mathcal{H}} \quad \mathcal{R}(h) \quad \text{subject to} \quad \gamma_i(h) \leq \epsilon_i, \quad \text{for } i = 1, \ldots, n \tag{1}$$

where $\mathcal{R}(h) = \mathbb{P}(h(X) \neq Y)$ is the classification error, which is defined as the probability that the prediction of the model $h(X)$ does not correspond to the true label Y. There are n constraints, each represented by $\gamma_i(h)$, which represents a fairness constraint expressed as a linear condition, with a threshold ϵ_i. These fairness constraints are central to the learning problem formulation, and we will discuss them in the next part.

3.1. Fairness Constraints

In the context of fairness constraints, the literature has proposed several group fairness constraints [4, 18, 19], each implementing different definitions of fairness between groups defined by the sensitive attribute A (see Section 2). Here, we give two widely used definitions of group fairness that can be generalised to both binary classification and multi-class classification settings [23].

Definition 1 (Demographic Parity)**.** *A classifier h is said to satisfy demographic parity if the probability of making a positive prediction is the same between all groups defined by the protected attribute A. In mathematical terms, it can be expressed as:*

$$\mathbb{P}(h(X) = y_p \mid A = a) = \mathbb{P}(h(X) = y_p), \quad \textit{for all } a \in \mathcal{A}, \tag{2}$$

where $y_p \in \{0, 1, \ldots, K\}$ denotes the positive or favorable class label.

Definition 2 (Equalized Odds)**.** *A classifier h meets the equalized odds fairness definition when the probability of predicting the favourable class label is the same across all groups determined by the protected attribute A, given the true label Y. In mathematical terms, this condition means:*

$$\mathbb{P}(h(X) = y_p \mid Y = y, A = a) = \mathbb{P}(h(X) = y_p \mid Y = y),\ \forall y \in \mathcal{Y},\ a \in \mathcal{A}, \quad (3)$$

In this equation, $y_p \in \{0, 1, \ldots, K\}$ refers to the positive or favorable class label, while $y \in \{0, 1, \ldots, K\}$ is the value of the true label.

To get back to the binary context, we only need to think about a label space $\mathcal{Y} = \{0, 1\}$ and treat $y_p = 1$ as the positive label. Then, the definitions above just turn into the usual binary forms that are often used in fairness studies [4].

Fairness constraints are often reworded with expectations to make them easier to integrate into optimisation problems. For binary case, where the label space is $\mathcal{Y} = \{0, 1\}$ and the classifier output is $h(X) \in \{0, 1\}$, Agarwal et al. [32] showed that definitions such as Demographic Parity and Equalized Odds can be written as linear constraints using expected values. This is because the chance of guessing the positive class $y_p = 1$ can be written as:

$$\mathbb{P}(h(X) = y_p) = \mathbb{E}[h(X)]. \quad (4)$$

where $\mathbb{E}[h(X)]$ is the expected value of the classifier output $h(X)$.

In case of a fair classifier, the expected value of $h(X)$ should be the same regardless of the value of the sensitive features. This leads to simplifying the expressions of fairness constraints. For example, the Demographic Parity constraint becomes

$$\mathbb{E}[h(X) \mid A = a] = \mathbb{E}[h(X)], \quad \forall a \in \mathcal{A}, \quad (5)$$

while the Equalized Odds condition becomes

$$\mathbb{E}[h(X) \mid A = a, Y = y] = \mathbb{E}[h(X) \mid Y = y], \quad \forall a \in \mathcal{A},\ y \in \mathcal{Y}. \quad (6)$$

This idea can also extend to the multi-class situation, where $h(X) \in \{0, 1, \ldots, K\}$. We apply indicator functions to separate the prediction of a specific class $y_p \in \mathcal{Y}$. In this instance, the chance of predicting y_p becomes

$$\mathbb{P}(h(X) = y_p) = \mathbb{E}[\mathbf{1}_{\{h(X)=y_p\}}], \tag{7}$$

with the indicator function defined as

$$\mathbf{1}_{\{h(X)=y_p\}} = \begin{cases} 1 & \text{if } h(X) = y_p, \\ 0 & \text{otherwise.} \end{cases}$$

Using this formulation, the fairness constraints can be expressed in a unified expectation-based form that applies to both binary and multi-class settings. The Demographic Parity constraint is then

$$\mathbb{E}[\mathbf{1}_{\{h(X)=y_p\}} \mid A = a] = \mathbb{E}[\mathbf{1}_{\{h(X)=y_p\}}] \quad \forall a \in \mathcal{A}, \tag{8}$$

while Equalized Odds become

$$\mathbb{E}[\mathbf{1}_{\{h(X)=y_p\}} \mid A = a, Y = y] = \mathbb{E}[\mathbf{1}_{\{h(X)=y_p\}} \mid Y = y] \quad \forall a \in \mathcal{A},\ y \in \mathcal{Y}. \tag{9}$$

These fairness notions can be reformulated in a structured manner that is compatible with linear optimization.

3.2. Fairness Constraints as Linear Moment Conditions

To make the fairness constraints more flexible and suitable for numerical optimisation, they are relaxed into linear inequalities of the classifier moments that take the form:

$$\gamma_i(h) = \sum_{k=1}^{m} M_{ik}\, \mu_j(h) \leq \epsilon_i, \quad i = 1, \ldots, n. \tag{10}$$

Here, M_{ij} are the entries of a matrix $M \in \mathbb{R}^{n \times m}$ that defines how each moment contributes to each constraint, where m is the number of moments and n is the number of constraints. The term ϵ_i denotes the upper bound for the i-th constraint, and $\mu_j(h)$ is the j-th moment of the classifier h, given by:

$$\mu_j(h) = \mathbb{E}[g_j(X, A, Y, h(X)) \mid E_j], \quad j = 1, \ldots, m, \tag{11}$$

In this formula, the function $g_j \colon \mathcal{X} \times \mathcal{A} \times \mathcal{Y} \times \mathcal{Y} \to [0, 1]$ is a measurable function influenced by the predicted label $h(X)$. The event E_j is a set condition on the variables (X, A, Y), like $A = a$ or $A = a\ \&\ Y = y$, and it does not depend on the model.

To illustrate the concept, consider the case of Demographic Parity. To address fairness under this definition, for each group $a \in \mathcal{A}$, we define the following moment

$$\mu_a(h) = \mathbb{E}[\mathbf{1}_{\{h(X)=y_p\}} \mid A = a], \tag{12}$$

This moment captures the probability that the classifier predicts the favourable class y_p for individuals within group a. Moreover, we define the overall moment:

$$\mu_*(h) = \mathbb{E}[\mathbf{1}_{\{h(X)=y_p\}}], \tag{13}$$

corresponding to the unconditional selection rate of class y_p across the entire population.

A classifier is fair under the Demographic Parity definition if it provides the same expected value of $\mathbf{1}_{\{h(X)=y_p\}}$ regardless of the value of A. Therefore, each equality constraint can be expressed as

$$\mu_a(h) = \mu_*(h), \quad \forall a \in \mathcal{A}, \tag{14}$$

which can be equivalently rewritten as the pair of inequalities:

$$\mu_a(h) - \mu_*(h) \leq 0,$$
$$\mu_*(h) - \mu_a(h) \leq 0.$$

In the binary sensitive attribute case where $A \in \{0, 1\}$, the group $A = 0$ is our *unprivileged* group and $A = 1$ our *privileged* group. Hence, in this case, these further inequalities hold as

$$\mu_0(h) - \mu_*(h) \leq 0,$$
$$\mu_*(h) - \mu_0(h) \leq 0,$$
$$\mu_1(h) - \mu_*(h) \leq 0,$$
$$\mu_*(h) - \mu_1(h) \leq 0.$$

In the case of a biased classifier, we expect $\mu_1(h) > \mu_*(h) > \mu_0(h)$. To achieve Demographic Parity, we build a constraint system of the form:

$$M\,\mu(h) \leq \epsilon, \quad \text{with} \quad \boldsymbol{\epsilon} = \begin{bmatrix} 0 & \cdots & 0 \end{bmatrix}^\top \in \mathbb{R}^4, \quad \mu(h) = \begin{bmatrix} \mu_0(h) \\ \mu_1(h) \\ \mu_*(h) \end{bmatrix}, \text{ and}$$

$$M = \begin{bmatrix} 1 & 0 & -1 \\ -1 & 0 & 1 \\ 0 & 1 & -1 \\ 0 & -1 & 1 \end{bmatrix}.$$

Here, the coefficients M_{ij} are simply $+1$, -1, or 0 depending on whether the moment $\mu_j(h)$ appears with a positive sign, a negative sign, or not at all. For instance, the inequality $\mu_0(h) - \mu_*(h) \leq 0$ corresponds to the row $[1, 0, -1]$ in M, while $\mu_*(h) - \mu_0(h) \leq 0$ corresponds to $[-1, 0, 1]$.

Similarly, for Equalized Odds, we impose that the classifier's prediction is independent of the sensitive attribute A conditionally on the true label $Y = y$. This leads to defining separate moments for each group $a \in \mathcal{A}$ and each label $y \in \mathcal{Y}$, of the form:

$$\mu_{a,y}(h) = \mathbb{E}[\mathbf{1}_{\{h(X)=y_p\}} \mid A = a, Y = y],$$

which represent the group-wise true positive (or false positive) rates, depending on the value of y_p. We also define the corresponding average moment across all groups:

$$\mu_{*,y}(h) = \mathbb{E}[\mathbf{1}_{\{h(X)=y_p\}} \mid Y = y],$$

which captures the overall prediction rate for class y_p conditioned on the true label $Y = y$. Equalized Odds is satisfied when

$$\mu_{a,y}(h) = \mu_{*,y}(h) \quad \forall a \in \mathcal{A},\ y \in \mathcal{Y}.$$

As before, each equality can be expressed as two inequalities:

$$\mu_{a,y}(h) - \mu_{*,y}(h) \leq 0$$
$$\mu_{*,y}(h) - \mu_{a,y}(h) \leq 0$$

In the binary sensitive attribute case $A \in \{0, 1\}$, we may identify a group $A = 0$ as the *unprivileged* group and $A = 1$ as the *privileged* group. Therefore, in this case, we obtain the following set of inequalities for each true label $y_p \in \mathcal{Y}$:

$$\mu_{0,y_p}(h) - \mu_{*,y_p}(h) \leq 0$$
$$\mu_{*,y_p}(h) - \mu_{0,y_p}(h) \leq 0$$
$$\mu_{1,y_p}(h) - \mu_{*,y_p}(h) \leq 0$$

$$\mu_{*,y_p}(h) - \mu_{1,y_p}(h) \leq 0$$

This leads to a constraint system of the form:

$$M\,\mu(h) \leq \epsilon, \quad \text{with} \quad \boldsymbol{\epsilon} = \begin{bmatrix} 0 & \cdots & 0 \end{bmatrix}^\top \in \mathbb{R}^4, \quad \mu(h) = \begin{bmatrix} \mu_{0,y_p}(h) \\ \mu_{1,y_p}(h) \\ \mu_{*,y_p}(h) \end{bmatrix}, \text{ and}$$

$$M = \begin{bmatrix} 1 & 0 & -1 \\ -1 & 0 & 1 \\ 0 & 1 & -1 \\ 0 & -1 & 1 \end{bmatrix}.$$

It is possible to enforce multiple fairness definitions simultaneously by combining their respective constraint formulations. In particular, one may impose both Demographic Parity and Equalized Odds as joint conditions on the classifier. This results in the following pair of fairness constraints:

$$\mu_a(h) = \mu_*(h), \qquad \text{and} \qquad \mu_{a,y}(h) = \mu_{*,y}(h), \quad \forall a \in \mathcal{A},\ y \in \mathcal{Y}.$$

Therefore, it can be expressed as inequalities as

$$\mu_a(h) - \mu_*(h) \leq 0$$
$$\mu_*(h) - \mu_a(h) \leq 0$$
$$\mu_{a,y}(h) - \mu_{*,y}(h) \leq 0$$
$$\mu_{*,y}(h) - \mu_{a,y}(h) \leq 0.$$

When the binary sensitive attribute is in the form $A \in \{0,1\}$, this gives us an inequality system:

$$\mu_0(h) - \mu_*(h) \leq 0$$
$$\mu_*(h) - \mu_0(h) \leq 0$$
$$\mu_1(h) - \mu_*(h) \leq 0$$
$$\mu_*(h) - \mu_1(h) \leq 0$$
$$\mu_{0,y_p}(h) - \mu_{*,y_p}(h) \leq 0$$
$$\mu_{*,y_p}(h) - \mu_{0,y_p}(h) \leq 0$$
$$\mu_{1,y_p}(h) - \mu_{*,y_p}(h) \leq 0$$

$$\mu_{*,y_p}(h) - \mu_{1,y_p}(h) \leq 0$$

We can compactly express this constraint system as:

$$M\,\mu(h) \leq \epsilon, \quad \text{with} \quad \epsilon = \begin{bmatrix} 0 & \cdots & 0 \end{bmatrix}^\top \in \mathbb{R}^8,\ \mu(h) = \begin{bmatrix} \mu_0(h) \\ \mu_1(h) \\ \mu_*(h) \\ \mu_{0,y_p}(h) \\ \mu_{1,y_p}(h) \\ \mu_{*,y_p}(h) \end{bmatrix},\ \text{and}$$

$$M = \begin{bmatrix} 1 & 0 & -1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & -1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix}.$$

After defining the fairness constraints as linear inequalities over conditional moments, we can now move to the optimization procedure. We introduce a general form of the Exponentiated Gradient algorithm of Agarwal et al. [32], which was first made for binary classification tasks and for only one fairness constraint. Our updated version supports both binary and multi-class classification tasks in the presence of multiple fairness constraints.

3.3. Generalized Exponentiated Gradient (GEG)

In this part, we provide a detailed description of the Generalized Exponentiated Gradient (GEG) method, an in-processing bias mitigation algorithm aimed at achieving fairness both in binary and multi-class classification tasks under multiple fairness definitions.

The primary goal of our approach is to find a classifier that yields the highest possible fairness while still being effective in its predictions, as formalised in the optimisation problem 1. Since the hypothesis space $\mathcal{H}$ is not convex and the loss function is not continuous, the direct optimisation of this problem faces significant computational challenges, which may lead to convergence issues, meaning the absence of an optimal solution. To address these difficulties, we adopt the method of Agarwal et al. [32] and rewrite the problem in terms of random classifiers, given as distributions $Q \in \Delta(\mathcal{H})$ over the hypothesis class. This relaxation transforms the original optimization problem into a convex one, allowing us to utilize efficient convex optimization methods. Thus, there is a solution to this problem that can be written as

$$\min_{Q \in \Delta(\mathcal{H})} \mathcal{R}(Q) \quad \text{subject to} \quad \gamma_i(Q) \leq \epsilon_i, \quad \text{for } i = 1, \ldots, n, \tag{15}$$

where $\Delta(\mathcal{H})$ is the set of all probability distributions over $\mathcal{H}$, $\mathcal{R}(Q) = \sum_{h \in \mathcal{H}} Q(h)\, \mathcal{R}(h)$ is the expected classification error under the randomized classifier Q, and $\gamma_i(Q) = \sum_{h \in \mathcal{H}} Q(h)\, \gamma_i(h)$ is the i-th expected fairness moment.

In a real-world scenario, we have access only to a finite training set and not the whole distribution of (X, A, Y). Thus, we obtain estimates for the expectations by taking averages over the sample at hand, and we allow for a bit slack $\widehat{\epsilon}_i$ in the constraint violations. In addition, the random classifier Q can be expressed as a sparse distribution on a set of predictors learned during training.

This brings us to the following approximation problem:

$$\min_{Q \in \Delta(\mathcal{H})} \widehat{\mathcal{R}}(Q) \quad \text{subject to} \quad \widehat{\gamma}_i(Q) \leq \widehat{\epsilon}_i, \quad i = 1, \ldots, n, \tag{16}$$

where $\widehat{\mathcal{R}}(Q), \widehat{\gamma}_i(Q)$ are learnt over the training samples, the tolerance constants $\widehat{\epsilon}_i$ are allowed to adapt at every learning iteration.

To solve the optimization problem (16), we reformulate it as a saddle-point problem using a Lagrangian approach. This transformation enables the use of duality principles from convex optimization and supports the design of efficient iterative algorithms. Specifically, we define the Lagrangian function:

$$\mathcal{L}(Q, \boldsymbol{\lambda}) = \widehat{\mathcal{R}}(Q) + \sum_{i=1}^{n} \lambda_i \left(\widehat{\gamma}_i(Q) - \widehat{\epsilon}_i \right), \tag{17}$$

where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_n)$ are the non-negative dual variables (Lagrange multipliers) associated with the fairness constraints.

The optimization problem is thus transformed into the following min-max saddle-point problem:

$$\min_{Q \in \Delta(\mathcal{H})} \max_{\boldsymbol{\lambda} \geq 0} \mathcal{L}(Q, \boldsymbol{\lambda}). \tag{18}$$

As a way to make the optimization problem more stable and better behaved, we restrict the dual domain by adding an ℓ_1-norm constraint on $\boldsymbol{\lambda}$. The resulting saddle-point formulation is:

$$\min_{Q\in\Delta(\mathcal{H})} \max_{\boldsymbol{\lambda}\geq 0,\ \|\boldsymbol{\lambda}\|_1\leq B} \mathcal{L}(Q,\boldsymbol{\lambda}). \tag{19}$$

According to Sion's minimax theorem [34], a solution to this problem is guaranteed to exist, since $\mathcal{L}(Q,\boldsymbol{\lambda})$ is linear in both arguments and the domains of Q and $\boldsymbol{\lambda}$ are convex and compact (the dual compactness is ensured by the ℓ_1-norm bound).

To find this saddle point, we follow the strategy used by Agarwal et al. [32], which frames the problem as a zero-sum game between two players. The *learner*, who selects a randomized classifier $Q \in \Delta(\mathcal{H})$ to minimize the classification loss while satisfying fairness constraints, and the *auditor*, who updates the dual variables $\boldsymbol{\lambda}$ to maximize the Lagrangian by penalizing constraint violations.

At each iteration, the learner constructs a new classifier $h_t \in \mathcal{H}$ by solving a *cost-sensitive classification problem*. This problem is formulated by assigning to each training sample $(x_j, a_j, y_j)_{j=1}^N$ a signed weight $w_j^{(t)}$ that combines two key components: the classification objective and the fairness constraints. Formally, we define:

$$w_j^{(t)} = \gamma_j^{\text{error}} + \sum_{i=1}^{n} \lambda_i^{(t)} \cdot \gamma_{i,j}^{\text{fair}},$$

where $\gamma_j^{\text{error}} \in \{+1,-1\}$ encodes the misclassification cost with respect to the target class y_p, defined as:

$$\gamma_j^{\text{error}} = \begin{cases} +1 & \text{if } y_j \neq y_p, \\ -1 & \text{if } y_j = y_p, \end{cases}$$

and $\gamma_{i,j}^{\text{fair}} \equiv \gamma_i(h(x_j))$ denotes the individual (per-sample) contribution of observation j to the violation of the i-th fairness constraint. The scalar $\lambda_i^{(t)}$ represents the Lagrange multiplier associated with this constraint at iteration t.

The sign of $w_j^{(t)}$ determines the target label in the cost-sensitive classification. The adjusted label $\tilde{y}_j^{(t)}$ is then set as:

$$\tilde{y}_j^{(t)} = \begin{cases} y_p & \text{if } w_j^{(t)} > 0, \\ y_j & \text{otherwise.} \end{cases}$$

The learner then solves the following cost-sensitive optimization problem:

$$h_t = \arg\min_{h \in \mathcal{H}} \sum_j |w_j^{(t)}| \cdot \mathbf{1}_{\{h(x_j) \neq \tilde{y}_j^{(t)}\}}.$$

Then, the auditor updates the dual variables $\boldsymbol{\lambda}^{(t)}$ by computing:

$$\lambda_i^{(t)} = \frac{B \cdot \exp(\theta_i^{(t)})}{1 + \sum_{k=1}^{n} \exp(\theta_k^{(t)})}, \quad \text{for all } i = 1, \ldots, n.$$

In practice, the update of $\boldsymbol{\theta}^{(t)}$ may also involve a learning rate or smoothing strategy to stabilize optimization, as implemented in our algorithm. This formula ensures that $\boldsymbol{\lambda}^{(t)}$ lies in the scaled probability simplex of radius B, emphasizing the most violated constraints.
The process converges to an approximate saddle point $(Q^\star, \boldsymbol{\lambda}^\star)$, which represents a randomized classifier that achieves an optimal balance between predictive performance and fairness. The resulting distribution $Q^\star$ is sparse, supported on a small number of base classifiers h_t, and is normalized to form a valid probability distribution over the hypothesis class. The final weights $\boldsymbol{\lambda}^\star$ provide insight into the most influential fairness constraints.

The optimization process stops when the *duality gap* falls below a small threshold ν, indicating that the current solution is close to a saddle point. The duality gap is computed as the difference between the Lagrangian value of the best single classifier and that of the current mixture Q:

$$\mathrm{Gap}(Q, \boldsymbol{\lambda}) = \max_{h \in \mathcal{H}} \mathcal{L}(h, \boldsymbol{\lambda}) - \mathcal{L}(Q, \boldsymbol{\lambda}).$$

When this gap becomes sufficiently small, no further improvement is expected, and the optimization terminates.
This entire procedure is formally presented in Algorithm 1.

Algorithm 1 Generalized Exponentiated Gradient (GEG)

Require: Training data (X, A, Y) with $Y \in \{0, 1, \ldots, K\}$, positive class y_p
Hypothesis class $\mathcal{H}$, fairness constraints $\{\widehat{\gamma}_i\}_{i=1}^n$ with thresholds $\{\epsilon_i\}_{i=1}^n$
Parameters: learning rate $\eta > 0$, tolerance $\delta > 0$, max iterations T, duality gap threshold $\nu > 0$, minimum iterations $t_{\min} \in \mathbb{N}$.
Ensure: Randomized classifier $Q \in \Delta(\mathcal{H})$

Initialize dual variables: $\boldsymbol{\theta} \leftarrow \mathbf{0}$, count vector $Q \leftarrow \emptyset$, budget $B \leftarrow 1/\delta$
for $t = 1$ to T **do**
    Compute dual weights: $\lambda_i^{(t)} \leftarrow \dfrac{B \cdot \exp(\theta_i^{(t)})}{1 + \sum_{k=1}^n \exp(\theta_k^{(t)})}$
    for each training sample $j = 1$ to N **do**
        Compute signed weight: $w_j^{(t)} \leftarrow \gamma_j^{\text{error}} + \sum_i \lambda_i^{(t)} \cdot \gamma_{i,j}^{\text{fair}}$
        Adjust label: $\tilde{y}_j \leftarrow \begin{cases} y_p & \text{if } w_j^{(t)} > 0 \\ y_j & \text{otherwise} \end{cases}$
        Normalize weights: $w_j^{(t)} \leftarrow \dfrac{N \cdot |w_j^{(t)}|}{\sum_{k=1}^N |w_k^{(t)}|}$ for all j
    end for
    Train classifier h_t on $\{(x_j, \tilde{y}_j, w_j^{(t)})\}_{j=1}^N$
    Update count: $Q[h_t] \leftarrow Q[h_t] + 1$
    Compute constraint violations: $\widehat{\gamma}_i(h_t)$
    Update dual: $\theta_i^{(t+1)} \leftarrow \theta_i^{(t)} + \eta \cdot (\widehat{\gamma}_i(h_t) - \epsilon_i)$
    Compute current mixture: $Q_t(h) \leftarrow \dfrac{Q(h)}{\sum_{h'} Q(h')}$ for all h
    Compute duality gap: $\text{Gap}_t \leftarrow \max_{h \in \mathcal{H}} \mathcal{L}(h, \boldsymbol{\lambda}) - \mathcal{L}(Q_t, \boldsymbol{\lambda})$
    if $\text{Gap}_t < \nu$ and $t \geq t_{\min}$ **then**
        break
    end if
end for
Normalize final distribution: $Q(h) \leftarrow \dfrac{Q(h)}{\sum_{h'} Q(h')}$ for all $h \in \mathcal{H}$
return Q

3.4. Implementation Details

We implemented GEG in Python 3.9 by extending the EG implementation provided by the `Fairlearn` Python library [35]. In all the experiments reported in Section 4, we set η to 10^{-5} and δ to 0.05. We provide the implementation of GEG and the evaluation scripts online for public use and research [16].

4. Evaluation

In this section, we describe the empirical evaluation conducted to assess the effectiveness of GEG. Specifically, our evaluation is driven by the following research questions (RQ):

RQ$_1$ **Multi-class classification:** *To what extent is GEG able to mitigate bias while keeping a high prediction effectiveness in a multi-class classification context?*

This RQ acts as a "sanity check" and benchmarks the ability of GEG in mitigating bias while keeping high prediction effectiveness against a base-classifier in the multiclass classification context.

RQ$_2$ **Binary classification:** *To what extent is GEG able to mitigate bias while keeping a high prediction effectiveness in a binary classification context?*

In addition to the multiclass classification context, we benchmark GEG against a base classifier employed in the binary classification context.

RQ$_3$ **Baseline comparison:** *How does GEG compare against existing bias mitigation methods in the multi-class classification tasks?*

In this RQ, we benchmark GEG against the *Debiaser for Multiple Variables (DEMV)* pre-processing approach, which is, to the best of our knowledge, the main approach proposed for bias mitigation in the multi-class classification context [23].

RQ$_4$ **Different base classifiers:** *How does GEG perform under different base-classifiers?*

Finally, this RQ evaluate the extent in which GEG can be effectively employed with different base-classifiers in the multi-class classification context.

In the following, we describe in detail the datasets employed (Section 4.1), the metrics used in the evaluation (Section 4.3), and the overall evaluation process (Section 4.2).

Table 1: Employed Datasets

Name	**Sens. Attribute**	**Instances**	**Features**	**Classes**	**Class Distr.**
CMC [36]	*religion*	1473	10	3	*1*: 42.7% *2*: 22.6% *3*: 34.7%
Crime [37]	*race*	1,994	100	6	*100*: 23.8% *200*: 15.8% *300*: 21.2% *400*: 19.9% *500*: 19.3%
Drug [38]	*race*	1,885	15	3	*0*: 21.9% *1*: 25.1% *2*: 52.9%
Law [3]	*gender*	20,427	14	3	*0*: 41.6% *1*: 27.9% *2*: 31.1%
Obesity [39]	*age*	1,490	17	5	*0*: 19.3% *1*: 19.5% *2*: 19.4% *3*: 23.5% *4*: 18.2%
Park [40]	*age*	5,875	19	3	*0*: 30.0% *1*: 44.6% *2*: 24.9%
Wine [41]	*type*	6,438	13	4	*4*: 3.4% *5*: 34.1% *6*: 45.3% *7*: 17.2%
Adult [42]	*sex*	30,940	102	2	*0*: 75.7% *1*: 24.2%
COMPAS [8]	*race*	6,167	399	2	*0*: 54.4% *1*: 45.5%
German [43]	*sex*	1,000	59	2	*0*: 30% *1*: 70%

4.1. Datasets

Table 1 reports the list of datasets employed in our study. For each dataset, we report its name, the sensitive attribute as reported in the corresponding source paper, the number of instances and features, the number of possible classes to be predicted, and their distribution. The datasets have been selected based on their relevance, diversity, and adoption in previous fairness studies [23, 44, 45]. Specifically, to answer $\mathbf{RQ}_1$, $\mathbf{RQ}_3$, and $\mathbf{RQ}_4$, we employ the following seven multi-class datasets:

1. **Contraceptive Method Choice (CMC)** [36]. This dataset contains 1,473 instances and 10 features about the adoption of contraceptive methods by women in Indonesia. The sensitive feature is *religion* and the positive outcome is *2 (long-term use)*.

2. **Communities and Crime (Crime)** [37]. This dataset includes 1,994 instances and 100 features about the per-capita violent crimes in U.S. communities. The sensitive feature is *race* and the positive outcome is *100 (low rate of crimes)*.

3. **Drug Usage (Drug)** [38]. This dataset includes 1,885 instances and 15 features about the frequency of drug consumption. The sensitive attribute is *race* and the positive class is *0 (never use)*.

4. **Law School Admission (Law)** [3]. This dataset contains 20,427 samples and 14 features about admissions scores to a law school. The sensitive attribute is *gender* and the positive outcome is *2 (high admission score)*.

5. **Obesity Estimation (Obesity)** [39]. This dataset contains 1,490 instances and 17 features about patients' obesity estimation. The sensitive feature is *age* and the positive class is *0 (no obesity)*.

6. **Parkinson's Telemonitoring (Park)** [40]. This dataset includes 5,875 instances and 19 features about patients affected by Parkinson's disease, measured with the Unified Parkinson's Disease Rating Scale (UPDRS) classification. The sensitive attribute is *age* and the positive class is *0 (mild class)*.

7. **Wine Quality (Wine)** [41]. This dataset includes 6,438 instances and 13 features about wine quality classification. The sensitive feature is wine *type* and the positive outcome is *6 (high quality class)*.

We employ instead the following binary datasets to answer the $\mathbf{RQ}_2$ of our study:

1. **Adult Income (Adult)** [42]. This dataset comprises 30,940 instances and 102 features related to the income of people in the U.S. The sensitive attribute is *sex* and the positive outcome is *1 (high income class)*.

2. **ProPublic Recidivism (COMPAS)** [8]. This dataset contains 6,167 samples and 399 features (one-hot encoded) about the recidivism prediction of condemned people. The sensitive feature is *race* and the positive class is *0 (no recidivism)*.

3. **German Credit (German)** [43]. This dataset includes 1,000 instances and 59 features about the classification of people as *good* or *bad* credit risk. The sensitive attribute is *sex* and the positive outcome is *1 (good credit risk)*.

4.2. Benchmarks

To answer the $\mathbf{RQ}_1$ and $\mathbf{RQ}_2$ of our study, we compare the fairness and effectiveness of GEG with those of a Logistic Regression (LR) classifier. We have chosen this model because it has been successfully applied in previous fairness studies and in multi-class classification tasks [18, 23]. To ensure a fair comparison, the same LR model is used as a base-classifier for GEG. Specifically, concerning $\mathbf{RQ}_1$, we employ three versions of GEG, each one adopting a different fairness constraint during the optimisation process: one version uses a Statistical Parity constraint (GEG-SP), another version employs the Equalised Odds constraint (GEG-EO), and the last version uses a Combined Parity constraint, optimising for both SP and EO at the same time (GEG-CP). Instead, concerning $\mathbf{RQ}_2$, since the implementation of GEG-SP and GEG-EO for binary classification is equal to the already existing EG approach from Agarwal et al. [15], we consider these methods as additional baselines. Therefore, we compare these results with those of GEG-CP, which is our novel contribution for binary classification.

Concerning $\mathbf{RQ}_3$, we compare the three versions of GEG (i.e., GEG-SP, GEG-EO, and GEG-CP) with the Debiaser for Multiple Variables (DEMV) approach, which is, to the best of our knowledge, the main approach proposed for bias mitigation in multi-class classification [23]. It is a pre-processing method that balances the dataset such that all the sensitive groups are equally represented. As for the first two **RQs**, we employ an LR model as a base classifier.

Finally, for $\mathbf{RQ}_4$, we benchmark the three versions of GEG against a Random Forest (RF) and a Gradient Boosting (GB) classifier. The choice for these models is still driven by their adoption in previous fairness studies and multi-class classification tasks [18, 23]. As with the other **RQs**, to ensure a fair comparison, we use the same base classifiers for GEG.

For all ML models, we use their implementation available in the *scikit-learn* Python library, with their default hyperparameters [46]. For DEMV, we employ the implementation available in the paper [23] with its default hyperparameters.

4.3. Evaluation Metrics and Methods

4.3.1. Metrics

We employ a heterogeneous set of metrics to evaluate the fairness and effectiveness of the approaches analysed in each RQ.

Concerning effectiveness metrics, following previous studies [44, 47], we employ the following metrics:

- **Accuracy**. This metric is defined as the percentage of correct predictions over the total predictions of a model:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}(\hat{y}_i = y_i)$$

  where N is the total number of samples, $\hat{y}_i$ and y_i are the i-th true and predicted samples, and $\mathbf{1}(\hat{y}_i = y_i)$ is a function which is equal to 1 if the prediction is equal to the true label and 0 otherwise. It ranges from 0 to 1, where 1 is the highest score [48].

- **Macro Precision**. This metric is an adaptation of the *Precision* score for the multi-class classification context [49]. It is defined as the unweighted average of the class-wise *precision* score:

$$\text{Macro Precision} = \frac{1}{K}\sum_{i=1}^{K} Precision_k$$

  where K is the number of classes and $Precision_k$ is the ratio of correctly predicted k class over all k predictions [49]. It ranges from 0 to 1, where 1 is the highest score.

- **Macro Recall**. Like Macro Precision, this metric is an adaptation of the *Recall* score for multi-class classification [49]. It is defined It is defined as the unweighted average of the class-wise *recall* score:

$$\text{Macro Recall} = \frac{1}{K}\sum_{i=1}^{K} Recall_k$$

where K is the number of classes and $Recall_k$ is the ratio of instances of the k class identified by the model [49]. It ranges from 0 to 1, where 1 is the highest score.

- **Macro F1 Score**. This metric is defined as the harmonic mean between *Macro Precision* and *Macro Recall* [49]:

$$\text{Macro F1 Score} = 2 \times \frac{\text{Macro Precision} \times \text{Macro Recall}}{\text{Macro Precision} + \text{Macro Recall}}$$

it ranges from 0 to 1, where 1 is the best score.

Concerning fairness, we consider three widely adopted fairness definitions [23, 44, 18]:

- **Statistical Parity Difference (SPD)**. This metric implements the *Demographic Parity* fairness definition defined in Definition 1. It measures fairness as the difference in the probability of having the positive outcome (y_p) predicted, being in the privileged group or not [26]. It is defined as:

$$\text{SPD} = Pr(\hat{y} = y_p | A = 0) - Pr(\hat{y} = y_p | A = 1)$$

where $A = 0$ and $A = 1$ are the unprivileged and privileged groups, respectively. This metric ranges from -1 to +1, and the closer to 0, the fairer the model.

- **Equal Opportunity Difference (EOD)**. This metric assesses fairness as the difference in the probability of having the positive outcome predicted conditioned on the value of the true label, being in the privileged group or not [28]. It is defined as:

$$\text{EOD} = Pr(\hat{y} = y_p | y = y_p, A = 0) - Pr(\hat{y} = y_p | y = y_p, A = 1)$$

this metric ranges from -1 to +1, and the closer to 0, the fairer the model.

- **Average Odds Difference (AOD)**. This metric implements the *Equalized Odds* fairness definition shown in Definition 2. It measures fairness as the difference between true positive (TPR) and false positive (FPR) rates, concerning the positive outcome, for items being in the privileged and unprivileged groups. Formally, it is defined as:

$$\text{AOD} = \frac{1}{2}((FPR_{A=0} - FPR_{A=1}) + (TPR_{A=0} - TPR_{A=1}))$$

like the other fairness metrics, this one ranges from -1 to +1, where the closer to 0, the fairer the model.

Following previous works [47, 44, 23], we consider absolute values of SPD, EOD, and AOD to have a clearer understanding of the fairness improvement in a model.

4.3.2. Methods

To mitigate the risk of data selection bias, for each **RQ**, we perform a 10-fold cross-validation with shuffling. For each fold, we train the models on the training set and compute the fairness and effectiveness metrics on the testing set. To ensure a fair evaluation, we use the same splits for all approaches in each RQ by fixing the random seed. Additionally, when we evaluate DEMV for the $\mathbf{RQ}_3$, following the original paper [23], we apply the pre-processing approach only on the training set.

After training and testing the approaches, we report the mean and standard deviation of the metrics obtained. Moreover, we employ the non-parametric one-sided Wilcoxon signed-rank test to assess the statistical significance of the difference between the metrics obtained by baselines and GEG. The Wilcoxon test is a non-parametric test that verifies the null hypothesis that the median between two dependent samples is different [50]. Being non-parametric, it raises the bar for significance by making no assumptions about the underlying samples. Specifically, the null hypothesis we check is *"H_0 : The objective O obtained by GEG is not improved with respect to the baseline approach x"*. The alternative hypothesis is: *"H_1 : The objective O obtained by GEG is improved with respect to the baseline approach x"*. For effectiveness metrics "*improved*" means that the score obtained by GEG is higher than the baseline. On the contrary, for fairness metrics "*improved*" means that the score obtained by GEG is lower than the baseline. Following standards [51, 52], we set the confidence value to 0.05. Therefore, we reject the null hypothesis if the test's p-value is < 0.05.

Table 2: RQ_1: Results for Multi-Class Classification against an LR Baseline. Winning cases are highlighted in blue, losing cases are highlighted in orange. Best fairness scores are highlighted in **bold**.

	Approach	Acc	Prec	Rec	F1	SPD	EOD	AOD
CMC	LR	0.606±0.033	0.583±0.037	0.56±0.032	0.553±0.033	0.115±0.066	0.178±0.129	0.116±0.076
	GEG - SP	0.602±0.028	0.575±0.032	0.556±0.026	0.549±0.027	**0.015±0.057**	0.045±0.147	0.02±0.063
	GEG - EO	0.601±0.034	0.574±0.038	0.56±0.031	0.557±0.032	0.028±0.045	**0.01±0.171**	**0.007±0.076**
	GEG - CP	0.606±0.03	0.576±0.038	0.561±0.032	0.556±0.034	0.058±0.062	0.057±0.162	0.042±0.087
Crime	LR	0.474±0.037	0.434±0.046	0.452±0.032	0.427±0.037	0.382±0.062	0.266±0.288	0.247±0.146
	GEG - SP	0.406±0.04	0.403±0.035	0.397±0.038	0.392±0.035	**0.028±0.048**	**0.056±0.174**	0.072±0.07
	GEG - EO	0.333±0.077	0.41±0.107	0.307±0.081	0.246±0.11	0.128±0.111	0.074±0.234	**0.066±0.136**
	GEG - CP	0.355±0.031	0.373±0.031	0.333±0.029	0.315±0.033	0.087±0.065	0.118±0.213	**0.06±0.103**
Drug	LR	0.687±0.023	0.618±0.033	0.614±0.018	0.611±0.025	0.213±0.125	0.276±0.181	0.177±0.1
	GEG - SP	0.683±0.025	0.613±0.031	0.606±0.024	0.604±0.026	**0.021±0.093**	0.087±0.179	0.058±0.088
	GEG - EO	0.667±0.029	0.584±0.039	0.586±0.027	0.576±0.031	0.043±0.118	0.09±0.262	0.034±0.141
	GEG - CP	0.684±0.028	0.609±0.029	0.606±0.023	0.6±0.024	0.066±0.121	**0.03±0.164**	**0.012±0.101**
Law	LR	0.666±0.015	0.64±0.016	0.652±0.014	0.644±0.015	0.083±0.019	0.039±0.048	0.051±0.025
	GEG - SP	0.679±0.008	0.653±0.007	0.667±0.008	0.655±0.007	**0.011±0.022**	0.014±0.048	0.016±0.03
	GEG - EO	0.67±0.018	0.645±0.016	0.657±0.016	0.648±0.015	0.039±0.019	**0.003±0.038**	0.009±0.02
	GEG - CP	0.692±0.017	0.666±0.018	0.68±0.018	0.664±0.015	0.038±0.021	0.009±0.039	**0.002±0.019**
Obesity	LR	0.668±0.044	0.654±0.046	0.665±0.033	0.651±0.038	0.049±0.043	**0.012±0.119**	0.011±0.067
	GEG - SP	0.654±0.042	0.636±0.041	0.653±0.03	0.634±0.037	**0.002±0.063**	0.109±0.117	0.062±0.072
	GEG - EO	0.621±0.054	0.612±0.05	0.619±0.046	0.604±0.051	0.037±0.08	0.054±0.12	0.018±0.088
	GEG - CP	0.662±0.04	0.656±0.035	0.659±0.03	0.646±0.031	0.041±0.048	0.025±0.151	**0.007±0.084**
Park	LR	0.473±0.02	0.33±0.039	0.402±0.014	0.351±0.017	0.214±0.072	0.323±0.127	0.232±0.081
	GEG - SP	0.477±0.025	0.438±0.073	0.407±0.017	0.369±0.025	**0.004±0.055**	0.074±0.102	0.011±0.066
	GEG - EO	0.443±0.033	0.435±0.025	0.437±0.028	0.431±0.029	0.015±0.05	**0.014±0.068**	**0.007±0.051**
	GEG - CP	0.454±0.02	0.431±0.024	0.423±0.024	0.42±0.025	0.042±0.037	0.045±0.098	0.032±0.048
Wine	LR	0.454±0.019	0.246±0.062	0.259±0.006	0.201±0.012	0.115±0.046	0.105±0.062	0.115±0.048
	GEG - SP	0.455±0.014	0.281±0.095	0.265±0.008	0.22±0.013	0.009±0.027	0.008±0.029	0.006±0.026
	GEG - EO	0.439±0.012	0.232±0.078	0.251±0.007	0.177±0.023	**0.002±0.031**	0.018±0.045	0.004±0.034
	GEG - CP	0.45±0.014	0.26±0.046	0.254±0.006	0.182±0.015	**0.002±0.027**	**0.006±0.035**	**0.002±0.028**

5. Results

In the following, we report the results of our empirical evaluation. In each table, we report in blue the winning cases (i.e., Wilcoxon p-value < 0.05 with respect to the baseline(s)), while we highlight in orange the losing cases (i.e., Wilcoxon p-value > 0.95 with respect to the baseline(s)). Additionally, for each dataset analysed, we highlight the best fairness score (i.e., the one closest to zero) in **bold**.

5.1. RQ_1: Multi-class classification.

Table 2 reports the results of the comparison of the fairness and effectiveness obtained by the three versions of GEG and the baseline LR model.

From the table, we observe how all versions of GEG significantly improve the fairness of the base classifier under all datasets and fairness definitions considered. The only dataset in which we do not see a significant improvement under all fairness definitions is *Obesity*, where the bias of the base LR classifier is also low. Surprisingly, the improvement in fairness does not always come at the cost of reduced effectiveness. In fact, only in *Crime* and *Obesity* we observe a significant reduction in all effectiveness scores by specific versions of GEG. This reduction could be explained by the higher number of classes in these two datasets (6 and 5, respectively; see Table 1), which may make the overall prediction task more complex for the model. We also observe a reduction in effectiveness by the GEG-EO model for the *Drug* dataset. Nevertheless, even if statistically significant, the loss in effectiveness is not large, with a maximum loss in accuracy of 0.14 points concerning GEG-EO with the Crime dataset. Indeed, with the *Law*, *Park*, and, partially, *Wine* datasets, we observe a statistically significant improvement also in effectiveness scores compared with the base LR classifier.

Finally, from the fairness scores in Table 2, we do not observe a clear winner among the three versions of GEG employed. This means that all three versions are effective in bias mitigation under all fairness definitions analysed. Notably, all versions of GEG achieve statistically significantly better results also under the AOD fairness definition, which is not directly optimised by the model.

> **Answer to RQ$_1$:** GEG significantly improves the fairness of an LR classifier under multiple multi-class datasets and fairness definitions. The improvement in fairness achieved by GEG does not come with a high cost in effectiveness. Indeed, the effectiveness in predictions obtained by GEG is even higher than the LR model in 3 out of 7 datasets.

5.2. RQ$_2$: Binary Classification

Table 3 reports the fairness and effectiveness achieved by GEG for binary classification. We recall that, in this context, our novel contribution is the extension of the original EG approach from Agarwal et al. [15] with the CP constraint (GEG-CP in Table 3).

From the table, we observe that GEG-CP is the approach achieving the best fairness results in 8 out of 9 cases analysed (89%). Notably, GEG-CP is also the only approach achieving statistically better results under AOD with the German dataset. However, this improvement in fairness comes at the cost of reduced effectiveness (especially Recall and F1 Score). This result suggests that GEG-CP tends to produce fewer positive outcomes across all groups.

Table 3: RQ_2: Results for Binary Classification against an LR model and the base EG approach from Agarwal et al. Winning cases are highlighted in blue, losing cases are highlighted in orange. Best fairness values are highlighted in **bold**.

	Approach	Acc	Prec	Rec	F1	SPD	EOD	AOD
Adult	LR	0.827±0.007	0.772±0.012	0.727±0.01	0.744±0.01	0.179±0.016	0.154±0.048	0.119±0.026
	EG - SP	0.825±0.006	0.774±0.012	0.711±0.008	0.732±0.009	0.077±0.016	0.13±0.071	0.058±0.036
	EG - EO	0.828±0.004	0.776±0.009	0.725±0.005	0.744±0.005	0.145±0.018	0.051±0.075	0.055±0.038
	GEG - CP	0.769±0.005	0.739±0.025	0.535±0.007	0.506±0.013	**0.008±0.007**	**0.012±0.031**	**0.009±0.017**
Compas	LR	0.675±0.021	0.675±0.02	0.666±0.02	0.666±0.021	0.174±0.04	0.102±0.031	0.15±0.042
	EG - SP	0.675±0.022	0.675±0.021	0.666±0.021	0.666±0.022	0.049±0.071	**0.014±0.069**	0.023±0.068
	EG - EO	0.669±0.018	0.669±0.018	0.66±0.017	0.66±0.017	0.031±0.055	0.026±0.041	**0.005±0.058**
	GEG - CP	0.611±0.031	0.695±0.028	0.578±0.025	0.52±0.049	**0.001±0.072**	**0.014±0.051**	0.019±0.073
German	LR	0.745±0.044	0.693±0.063	0.661±0.058	0.669±0.06	0.21±0.108	0.182±0.17	0.17±0.13
	EG - SP	0.745±0.062	0.694±0.088	0.66±0.071	0.668±0.077	0.067±0.147	0.075±0.158	0.037±0.178
	EG - EO	0.748±0.056	0.698±0.079	0.664±0.067	0.672±0.072	0.1±0.169	0.088±0.189	0.059±0.194
	GEG - CP	0.708±0.051	0.613±0.193	0.528±0.029	0.479±0.056	**0.022±0.047**	**0.016±0.028**	**0.029±0.058**

Therefore, practitioners can choose to adopt GEG-CP in use cases where a reduction in positive outcomes is acceptable to achieve higher fairness (e.g., in use cases protected by specific regulations).

> **Answer to RQ_2:** In the binary classification context, GEG-CP achieves the best fairness reduction in 88% of the cases analysed compared to baselines. However, this improvement comes at the cost of a reduced ability of the model to deliver positive outcomes.

5.3. RQ_3: Baseline Comparison

Table 4 shows the results of the comparison between the three versions of GEG and the DEMV baseline for multi-class classification.

We observe how GEG achieves the best fairness scores in 20 out of 21 cases analysed (95%). The improvement achieved by GEG is also statistically significant in the *Crime*, *Law*, *Park*, and *Wine* datasets. Additionally, the effectiveness achieved by GEG is mostly comparable with that achieved by DEMV. Specifically, we observe how GEG-EO and, partially, GEG-CP tend to provide statistically significantly lower Precision, Recall and, consequently, F1 Score, under specific datasets (namely *Crime*, *Drug*, *Obesity*, and, partially, *Wine*). However, this decrease is observed primarily in datasets with a high number of classes or a high class imbalance (see Table 1). In fact, this reduction in Precision and Recall does not impact Accuracy in a statistically significant manner. On the contrary, GEG-SP provides consistent effectiveness across all datasets, whereas all versions of GEG show a con-

Table 4: RQ_3: Comparison with the DEMV pre-processing approach for multi-class classification. Winning cases are highlighted in blue, losing cases are highlighted in orange. Best fairness values are highlighted in **bold**.

	Approach	Acc	Prec	Rec	F1	SPD	EOD	AOD
CMC	DEMV	0.601±0.024	0.581±0.036	0.55±0.018	0.542±0.02	0.056±0.044	0.053±0.163	0.036±0.065
	GEG - SP	0.602±0.028	0.575±0.032	0.556±0.026	0.549±0.027	**0.015±0.057**	0.045±0.147	0.02±0.063
	GEG - EO	0.601±0.034	0.574±0.038	0.56±0.031	0.557±0.032	0.028±0.045	**0.01±0.171**	**0.007±0.076**
	GEG - CP	0.606±0.03	0.576±0.038	0.561±0.032	0.556±0.034	0.058±0.062	0.057±0.162	0.042±0.087
Crime	DEMV	0.451±0.032	0.406±0.046	0.429±0.021	0.398±0.031	0.324±0.053	0.224±0.268	0.199±0.121
	GEG - SP	0.406±0.04	0.403±0.035	0.397±0.038	0.392±0.035	**0.028±0.048**	**0.056±0.174**	0.072±0.07
	GEG - EO	0.333±0.077	0.41±0.107	0.307±0.081	0.246±0.11	0.128±0.111	0.074±0.234	**0.066±0.136**
	GEG - CP	0.355±0.031	0.373±0.031	0.333±0.029	0.315±0.033	0.087±0.065	0.118±0.213	**0.06±0.103**
Drug	DEMV	0.687±0.029	0.624±0.039	0.61±0.028	0.612±0.031	0.098±0.1	**0.039±0.157**	0.033±0.093
	GEG - SP	0.683±0.025	0.613±0.031	0.606±0.024	0.604±0.026	**0.021±0.093**	0.087±0.179	0.058±0.088
	GEG - EO	0.667±0.029	0.584±0.039	0.586±0.027	0.576±0.031	0.043±0.118	0.09±0.262	0.034±0.141
	GEG - CP	0.684±0.028	0.609±0.029	0.606±0.023	0.6±0.024	0.066±0.121	**0.03±0.164**	**0.012±0.101**
Law	DEMV	0.669±0.019	0.645±0.019	0.658±0.019	0.649±0.019	0.063±0.01	0.02±0.045	0.03±0.023
	GEG - SP	0.679±0.008	0.653±0.007	0.667±0.008	0.655±0.007	**0.011±0.022**	0.014±0.048	0.016±0.03
	GEG - EO	0.67±0.018	0.645±0.016	0.657±0.016	0.648±0.015	0.039±0.019	**0.003±0.038**	0.009±0.02
	GEG - CP	0.692±0.017	0.666±0.018	0.68±0.018	0.664±0.015	0.038±0.021	0.009±0.039	**0.002±0.019**
Obesity	DEMV	0.661±0.044	0.655±0.041	0.66±0.036	0.648±0.039	0.05±0.047	**0.014±0.159**	0.014±0.087
	GEG - SP	0.654±0.042	0.636±0.041	0.653±0.03	0.634±0.037	**0.002±0.063**	0.109±0.117	0.062±0.072
	GEG - EO	0.621±0.054	0.612±0.05	0.619±0.046	0.604±0.051	0.037±0.08	0.054±0.12	0.018±0.088
	GEG - CP	0.662±0.04	0.656±0.035	0.659±0.03	0.646±0.031	0.041±0.048	0.025±0.151	**0.007±0.084**
Park	DEMV	0.466±0.018	0.414±0.088	0.394±0.016	0.349±0.022	0.155±0.063	0.224±0.124	0.163±0.072
	GEG - SP	0.477±0.025	0.438±0.073	0.407±0.017	0.369±0.025	**0.004±0.055**	0.074±0.102	0.011±0.066
	GEG - EO	0.443±0.033	0.435±0.025	0.437±0.028	0.431±0.029	0.015±0.05	**0.014±0.068**	**0.007±0.051**
	GEG - CP	0.454±0.02	0.431±0.024	0.423±0.024	0.42±0.025	0.042±0.037	0.045±0.098	0.032±0.048
Wine	DEMV	0.453±0.015	0.26±0.084	0.259±0.006	0.204±0.011	0.143±0.057	0.132±0.068	0.143±0.058
	GEG - SP	0.455±0.014	0.281±0.095	0.265±0.008	0.22±0.013	0.009±0.027	0.008±0.029	0.006±0.026
	GEG - EO	0.439±0.012	0.232±0.078	0.251±0.007	0.177±0.023	**0.002±0.031**	0.018±0.045	0.004±0.034
	GEG - CP	0.45±0.014	0.26±0.046	0.254±0.006	0.182±0.015	**0.002±0.027**	**0.006±0.035**	**0.002±0.028**

sistently, and even statistically significantly, larger effectiveness in datasets with a more balanced label distribution.

> **Answer to RQ_3:** GEG achieves better fairness scores than DEMV in 95% of the cases analysed. Regarding effectiveness, GEG-SP provides consistent, or even statistically significantly better scores than DEMV. On the contrary, GEG-EO and, to a lesser extent, GEG-CP may struggle more with datasets with a high number of classes or high class imbalance.

5.4. RQ_4: Different Base Classifiers

Table 5 presents the results of applying GEG with an RF classifier. We observe that the RF classifier generally demonstrates higher predictive performance compared to the LR model. Concerning fairness, in all use cases,

Table 5: RQ$_4$: Results obtained with RF base classifier. Winning cases are highlighted in blue, losing cases are highlighted in orange. Best fairness scores are highlighted in **bold**.

	Approach	Acc	Prec	Rec	F1	SPD	EOD	AOD
CMC	RF	0.98±0.015	0.979±0.016	0.977±0.017	0.978±0.017	0.136±0.08	0.022±0.087	**0.012±0.045**
	GEG - SP	0.868±0.022	0.869±0.02	0.882±0.02	0.86±0.025	**0.029±0.092**	0.015±0.083	0.058±0.068
	GEG - EO	0.984±0.009	0.985±0.008	0.981±0.01	0.983±0.009	0.141±0.084	0.034±0.078	0.019±0.04
	GEG - CP	0.794±0.041	0.833±0.022	0.833±0.032	0.79±0.043	0.04±0.16	**0.01±0.062**	0.036±0.083
Crime	RF	0.499±0.031	0.471±0.03	0.478±0.025	0.466±0.027	0.413±0.069	0.483±0.232	0.355±0.14
	GEG - SP	0.437±0.04	0.441±0.033	0.411±0.022	0.395±0.03	**0.155±0.055**	**0.307±0.244**	**0.139±0.124**
	GEG - EO	0.507±0.036	0.473±0.036	0.484±0.029	0.47±0.034	0.424±0.061	0.497±0.194	0.367±0.117
	GEG - CP	0.415±0.04	0.442±0.032	0.386±0.03	0.364±0.036	0.19±0.046	0.379±0.226	0.197±0.113
Drug	RF	0.676±0.032	0.606±0.033	0.597±0.027	0.597±0.028	0.143±0.125	0.177±0.194	0.108±0.128
	GEG - SP	0.598±0.034	0.544±0.034	0.547±0.032	0.531±0.028	**0.029±0.11**	0.132±0.201	0.024±0.108
	GEG - EO	0.676±0.019	0.605±0.029	0.598±0.025	0.597±0.025	0.179±0.099	0.211±0.141	0.142±0.087
	GEG - CP	0.647±0.022	0.574±0.034	0.58±0.02	0.544±0.034	0.046±0.121	**0.101±0.175**	**0.011±0.104**
Law	RF	0.969±0.006	0.971±0.005	0.97±0.006	0.97±0.005	0.162±0.024	0.151±0.053	0.079±0.026
	GEG - SP	0.951±0.007	0.952±0.006	0.954±0.006	0.953±0.006	**0.027±0.021**	0.126±0.048	**0.015±0.023**
	GEG - EO	0.97±0.005	0.972±0.004	0.971±0.005	0.971±0.005	0.159±0.024	0.14±0.042	0.073±0.022
	GEG - CP	0.918±0.007	0.922±0.006	0.929±0.006	0.923±0.006	0.087±0.026	**0.125±0.04**	0.027±0.019
Obesity	RF	0.929±0.017	0.934±0.013	0.928±0.018	0.927±0.016	0.064±0.058	0.018±0.055	**0.0±0.032**
	GEG - SP	0.901±0.031	0.916±0.024	0.9±0.031	0.9±0.031	**0.012±0.07**	0.025±0.055	0.034±0.045
	GEG - EO	0.931±0.013	0.935±0.011	0.929±0.013	0.929±0.012	0.05±0.049	0.019±0.035	0.009±0.028
	GEG - CP	0.926±0.016	0.932±0.015	0.924±0.015	0.924±0.015	0.061±0.044	**0.002±0.062**	0.006±0.036
Park	RF	0.853±0.014	0.863±0.011	0.851±0.016	0.856±0.013	0.013±0.035	0.205±0.076	0.084±0.043
	GEG - SP	0.815±0.012	0.824±0.01	0.814±0.014	0.817±0.012	0.058±0.036	0.218±0.062	0.138±0.039
	GEG - EO	0.852±0.012	0.863±0.01	0.849±0.013	0.855±0.011	0.014±0.036	**0.204±0.07**	**0.083±0.042**
	GEG - CP	0.855±0.01	0.865±0.007	0.853±0.013	0.858±0.01	**0.008±0.047**	0.211±0.069	0.09±0.043
Wine	RF	0.709±0.015	0.77±0.046	0.556±0.023	0.589±0.031	0.122±0.036	0.093±0.037	0.093±0.036
	GEG - SP	0.708±0.012	0.731±0.086	0.548±0.025	0.579±0.036	**0.067±0.041**	**0.064±0.031**	**0.041±0.041**
	GEG - EO	0.709±0.015	0.755±0.083	0.55±0.025	0.581±0.033	0.119±0.051	0.092±0.053	0.091±0.049
	GEG - CP	0.707±0.012	0.726±0.077	0.548±0.02	0.578±0.03	0.121±0.044	0.091±0.039	0.093±0.044

at least one version of GEG achieves higher fairness scores than the RF baselines. However, it is important to note that this increase in fairness often results in reduced effectiveness. This decline in effectiveness can be attributed to the high effectiveness of the baseline RF classifier, which leads to a systematic trade-off: to enhance fairness, the overall effectiveness is typically lowered [18, 53].

Similar results are observed when employing a GB base classifier, as shown in Table 6. Notably, the baseline GB model emerges as the most effective and fair classifier among the three models analysed for multi-class classification. However, GEG still mitigates bias significantly when the base classifier's bias is relatively high (as in the *Crime* dataset). Finally, we observe some cases in which the bias of GEG exceeds that of the baseline model under the AOD definition. These issues can be explained by the low bias of

Table 6: RQ$_4$: Results obtained with GB base classifier. Winning cases are highlighted in blue, losing cases are highlighted in orange. Best fairness scores are highlighted in **bold**.

	Approach	Acc	Prec	Rec	F1	SPD	EOD	AOD
CMC	GB	0.996±0.005	0.995±0.006	0.996±0.005	0.995±0.005	0.134±0.104	0.011±0.017	**0.004±0.008**
	GEG - SP	0.869±0.033	0.877±0.027	0.886±0.029	0.861±0.035	**0.014±0.113**	**0.004±0.012**	0.09±0.027
	GEG - EO	0.996±0.005	0.995±0.006	0.996±0.005	0.995±0.005	0.134±0.104	0.011±0.017	**0.004±0.008**
	GEG - CP	0.801±0.062	0.844±0.033	0.841±0.042	0.798±0.062	0.04±0.134	0.007±0.014	0.044±0.071
Crime	GB	0.492±0.046	0.468±0.045	0.472±0.038	0.463±0.041	0.417±0.042	0.534±0.216	0.38±0.101
	GEG - SP	0.425±0.044	0.43±0.046	0.408±0.04	0.404±0.041	**0.04±0.058**	**0.127±0.266**	**0.009±0.136**
	GEG - EO	0.481±0.038	0.454±0.037	0.462±0.034	0.452±0.035	0.34±0.041	0.362±0.18	0.265±0.087
	GEG - CP	0.364±0.047	0.4±0.06	0.339±0.04	0.32±0.048	0.179±0.085	0.342±0.265	0.208±0.124
Drug	GB	0.676±0.031	0.606±0.032	0.602±0.027	0.6±0.029	0.167±0.13	0.18±0.156	0.127±0.115
	GEG - SP	0.674±0.021	0.598±0.029	0.593±0.023	0.592±0.024	**0.012±0.065**	0.085±0.154	**0.05±0.056**
	GEG - EO	0.681±0.034	0.613±0.039	0.608±0.032	0.605±0.036	0.113±0.102	0.108±0.156	0.065±0.094
	GEG - CP	0.663±0.038	0.588±0.038	0.587±0.033	0.583±0.034	0.029±0.104	**0.06±0.157**	0.034±0.085
Law	GB	1.0±0.0	1.0±0.0	1.0±0.0	1.0±0.0	0.134±0.024	**0.0±0.001**	**0.0±0.001**
	GEG - SP	0.979±0.004	0.979±0.004	0.981±0.003	0.979±0.004	**0.004±0.029**	**0.0±0.001**	0.081±0.012
	GEG - EO	1.0±0.0	1.0±0.0	1.0±0.0	1.0±0.0	0.134±0.024	**0.0±0.001**	**0.0±0.001**
	GEG - CP	0.944±0.007	0.949±0.006	0.954±0.006	0.948±0.007	0.056±0.031	**0.0±0.001**	0.044±0.016
Obesity	GB	0.95±0.014	0.948±0.014	0.948±0.014	0.947±0.014	0.051±0.058	0.015±0.099	**0.007±0.055**
	GEG - SP	0.929±0.022	0.931±0.018	0.927±0.02	0.927±0.021	**0.01±0.072**	**0.007±0.082**	0.028±0.057
	GEG - EO	0.95±0.014	0.948±0.014	0.948±0.014	0.947±0.014	0.051±0.058	0.015±0.099	0**0.007±0.055**
	GEG - CP	0.95±0.014	0.948±0.014	0.948±0.014	0.947±0.014	0.051±0.058	0.015±0.099	**0.007±0.055**
Park	GB	0.867±0.013	0.88±0.01	0.864±0.014	0.87±0.012	0.031±0.04	0.166±0.064	0.061±0.037
	GEG - SP	0.87±0.013	0.885±0.01	0.866±0.014	0.873±0.012	**0.011±0.042**	0.188±0.068	0.082±0.039
	GEG - EO	0.667±0.023	0.734±0.018	0.651±0.021	0.646±0.027	0.073±0.044	**0.123±0.063**	0.013±0.039
	GEG - CP	0.849±0.018	0.859±0.02	0.852±0.013	0.854±0.016	0.062±0.054	0.134±0.061	**0.029±0.042**
Wine	GB	0.606±0.015	0.566±0.043	0.453±0.02	0.475±0.026	0.116±0.038	0.055±0.067	0.092±0.035
	GEG - SP	0.6±0.015	0.559±0.049	0.452±0.015	0.475±0.022	**0.006±0.046**	0.032±0.059	**0.014±0.043**
	GEG - EO	0.6±0.016	0.554±0.042	0.445±0.019	0.466±0.025	0.041±0.039	**0.001±0.054**	0.02±0.032
	GEG - CP	0.606±0.019	0.564±0.042	0.453±0.015	0.475±0.02	0.064±0.048	0.017±0.057	0.043±0.042

the baseline classifier. Therefore, in these cases, applying a bias mitigation approach may not be needed. Nevertheless, even when higher than the baseline, the bias achieved by GEG is still low and not alarming (all values are < 0.1).

> **Answer to RQ$_4$:** GEG is effective in bias mitigation even when more complex base classifiers are employed, especially when the bias of the base classifier is relatively high.

5.5. Practical Insights

From our empirical analysis, we can draw the following main insights and recommendations on using GEG:

- GEG is effective in bias mitigation for multi-class classification regardless of the base classifier employed.

- When employing an LR classifier for multi-class classification tasks, adopting GEG in use cases where the number of classes to predict is ≤ 4 can also increase the effectiveness of the model.

- When adopting more complex classifiers such as RF or GB for multi-class classification, GEG is still effective in bias mitigation, but it may decrease the prediction's effectiveness. Nevertheless, this decrease may be systemic to achieve higher fairness with highly effective classifiers.

- GB emerged as the most effective and fair model for multi-class classification. Nevertheless, we show that GEG is effective at mitigating bias when the bias of the base GB model is relatively high.

- Concerning binary classification, users can employ GEG-CP in use cases where higher fairness is more relevant than having more positive outcomes predicted (e.g., use cases protected by specific regulations [7]).

- We suggest adopting GEG instead of the pre-preprocessing DEMV approach to achieve higher fairness. Additionally, when applied to datasets with low class imbalance, GEG can achieve higher prediction effectiveness than DEMV.

6. Conclusion and Future Work

In this paper, we addressed the topic of bias mitigation in multi-class classification settings. We first formulate the problem of fair multi-class learning as a multi-objective optimisation problem under multiple linear fairness constraints. Next, we propose GEG, an in-processing bias mitigation method to solve this task. In particular, GEG extends the EG approach from Agarwal et al. [15] to the multi-class classification setting. In addition, GEG allows the optimisation of a classifier under multiple fairness constraints simultaneously. We perform an extensive evaluation of GEG against six baseline approaches across seven multi-class and three binary datasets, using four effectiveness metrics and three fairness definitions. Our evaluation shows that GEG is successful at mitigating bias without severely impacting the effectiveness of the predictions. Additionally, we draw a set of practical insights for practitioners on using GEG in real-world scenarios.

Future work can extend GEG by including additional fairness constraints. Additionally, GEG can be extended to address intersectional fairness scenarios, i.e., where sensitive groups are identified by the combination of two or more sensitive variables [47].

References

[1] S. Canali, V. Schiaffonati, A. Aliverti, Challenges and recommendations for wearable devices in digital health: Data quality, interoperability, health equity, fairness, PLOS Digital Health 1 (10) (2022) e0000104, publisher: Public Library of Science. doi:10.1371/journal.pdig.0000104.
URL https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000104

[2] N. Kozodoi, J. Jacob, S. Lessmann, Fairness in credit scoring: Assessment, implementation and profit implications, European Journal of Operational Research 297 (3) (2022) 1083–1094, publisher: North-Holland. doi:10.1016/J.EJOR.2021.06.023.

[3] K. A. Austin, C. M. Christopher, D. Dickerson, Will i pass the bar exam: Predicting student success using lsat scores and law school performance, HofstrA l. rev. 45 (2016) 753.

[4] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A Survey on Bias and Fairness in Machine Learning, ACM Computing Surveys 54 (6) (2021) 1–35. doi:10.1145/3457607.

[5] S. Caton, C. Haas, Fairness in Machine Learning: A Survey, ACM Computing SurveysJust Accepted (2023). doi:10.1145/3616865.
URL https://dl.acm.org/doi/10.1145/3616865

[6] Z. Chen, J. M. Zhang, M. Hort, F. Sarro, M. Harman, Fairness Testing: A Comprehensive Survey and Analysis of Trends, arXiv:2207.10223 [cs] (Aug. 2022).
URL http://arxiv.org/abs/2207.10223

[7] EU AI Act: first regulation on artificial intelligence | News | European Parliament (Aug. 2023).
URL https://www.europarl.europa.eu/news/

en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

[8] J. Angwin, J. Larson, S. Mattu, L. Kirchner, Machine bias, ProPublica, May 23 (2016) (2016) 139–159.

[9] A. Baskota, Y.-K. Ng, A graduate school recommendation system using the multi-class support vector machine and knn approaches, in: 2018 IEEE International Conference on Information Reuse and Integration (IRI), IEEE, 2018, pp. 277–284.

[10] N. Yanes, A. M. Mostafa, M. Ezz, S. N. Almuayqil, A machine learning-based recommender system for improving students learning experiences, IEEE Access 8 (2020) 201218–201235.

[11] M. S. Suchithra, M. L. Pai, Improving the Performance of Sigmoid Kernels in Multiclass SVM Using Optimization Techniques for Agricultural Fertilizer Recommendation System, in: I. Zelinka, R. Senkerik, G. Panda, P. S. Lekshmi Kanthan (Eds.), Soft Computing Systems, Communications in Computer and Information Science, Springer, Singapore, 2018, pp. 857–868. doi:10.1007/978-981-13-1936-5_87.

[12] L. Meenachi, S. Ramakrishnan, M. Sivaprakash, C. Thangaraj, S. Sethupathy, Multi Class Ensemble Classification for Crop Recommendation, in: 2022 International Conference on Inventive Computation Technologies (ICICT), 2022, pp. 1319–1324, iSSN: 2767-7788. doi:10.1109/ICICT54344.2022.9850561.

[13] J. Zhang, P. Cao, D. P. Gross, O. R. Zaiane, On the application of multi-class classification in physical therapy recommendation, Health Information Science and Systems 1 (1) (2013) 15. doi:10.1186/2047-2501-1-15.
URL https://doi.org/10.1186/2047-2501-1-15

[14] U. Nations, THE 17 GOALS | Sustainable Development (2015).
URL https://sdgs.un.org/goals

[15] A. Agarwal, A. Beygelzimer, M. Dudik, J. Langford, H. Wallach, A reductions approach to fair classification, in: J. Dy, A. Krause (Eds.), Proceedings of the 35th International Conference on Machine Learning,

Vol. 80 of Proceedings of Machine Learning Research, PMLR, 2018, pp. 60–69.
URL https://proceedings.mlr.press/v80/agarwal18a.html

[16] M. Boubekraoui, G. d'Aloisio, A. Di Marco, GEG Replication Package (2025).
URL https://github.com/giordanoDaloisio/GEG

[17] T. Speicher, H. Heidari, N. Grgic-Hlaca, K. P. Gummadi, A. Singla, A. Weller, M. B. Zafar, A unified approach to quantifying algorithmic unfairness: Measuring individual &group unfairness via inequality indices, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 2239–2248.

[18] M. Hort, Z. Chen, J. M. Zhang, M. Harman, F. Sarro, Bias mitigation for machine learning classifiers: A comprehensive survey, ACM Journal on Responsible Computing 1 (2) (2024) 1–52.

[19] G. d'Aloisio, C. Di Sipio, A. Di Marco, D. Di Ruscio, How fair are we? from conceptualization to automated assessment of fairness definitions, Software and Systems Modeling (2025) 1–27.

[20] M. Hort, J. M. Zhang, F. Sarro, M. Harman, Fairea: A model behaviour mutation approach to benchmarking bias mitigation methods, in: Proceedings of the 29th ACM joint meeting on European software engineering conference and symposium on the foundations of software engineering, 2021, pp. 994–1006.

[21] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, D. Roth, A comparative study of fairness-enhancing interventions in machine learning, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, ACM, 2019, pp. 329–338.

[22] F. Kamiran, T. Calders, Data preprocessing techniques for classification without discrimination, Knowledge and Information Systems 33 (1) (2012) 1–33. doi:10.1007/s10115-011-0463-8.

[23] G. d'Aloisio, A. D'Angelo, A. Di Marco, G. Stilo, Debiaser for Multiple Variables to enhance fairness in classification tasks, Information Processing & Management 60 (2) (2023) 103226.

doi:10.1016/j.ipm.2022.103226.
URL https://www.sciencedirect.com/science/article/pii/S0306457322003272

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

[25] S. Tizpaz-Niari, A. Kumar, G. Tan, A. Trivedi, Fairness-aware configuration of machine learning libraries, arXiv preprint arXiv:2202.06196 (2022).

[26] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. Zemel, Fairness through awareness, in: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, ITCS '12, Association for Computing Machinery, New York, NY, USA, 2012, pp. 214–226. doi:10.1145/2090236.2090255.

[27] Z. Chen, J. M. Zhang, F. Sarro, M. Harman, Maat: a novel ensemble approach to addressing fairness and performance bugs for machine learning software, in: Proceedings of the 30th ACM joint european software engineering conference and symposium on the foundations of software engineering, 2022, pp. 1122–1134.

[28] M. Hardt, E. Price, N. Srebro, Equality of opportunity in supervised learning, Advances in neural information processing systems 29 (2016) 3315–3323.

[29] F. Kamiran, T. Calders, M. Pechenizkiy, Discrimination aware decision tree learning, in: 2010 IEEE International Conference on Data Mining, IEEE, 2010, pp. 869–874.

[30] P. Putzel, S. Lee, Blackbox Post-Processing for Multiclass Fairness, arXiv:2201.04461 [cs]ArXiv: 2201.04461 (Jan. 2022).
URL http://arxiv.org/abs/2201.04461

[31] C. Denis, R. Elie, M. Hebiri, F. Hu, Fairness guarantees in multi-class classification with demographic parity, Journal of Machine Learning Research 25 (130) (2024) 1–46.

[32] 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.

[33] S. Radovanović, A. Petrović, B. Delibašić, M. Suknović, Enforcing fairness in logistic regression algorithm, in: 2020 International Conference on INnovations in Intelligent SysTems and Applications (INISTA), IEEE, 2020, pp. 1–7.

[34] M. Sion, On general minimax theorems, Pacific Journal of Mathematics 8 (1) (1958) 171–176.

[35] Fairlearn, Fairlearn documentation (2022).
URL `https://fairlearn.org/main/faq.html`

[36] T.-S. Lim, W.-Y. Loh, Y.-S. Shih, A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms, Machine learning 40 (3) (2000) 203–228.

[37] M. Redmond, A. Baveja, A data-driven software tool for enabling cooperative information sharing among police departments, European Journal of Operational Research 141 (3) (2002) 660–678.

[38] E. Fehrman, A. K. Muhammad, E. M. Mirkes, V. Egan, A. N. Gorban, The Five Factor Model of Personality and Evaluation of Drug Consumption Risk, in: F. Palumbo, A. Montanari, M. Vichi (Eds.), Data Science, Studies in Classification, Data Analysis, and Knowledge Organization, Springer International Publishing, Cham, 2017, pp. 231–242. `doi:10.1007/978-3-319-55723-6_18`.

[39] F. M. Palechor, A. d. l. H. Manotas, Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico, Data in Brief 25 (2019) 104344. `doi:10.1016/j.dib.2019.104344`.
URL `https://www.sciencedirect.com/science/article/pii/S2352340919306985`

[40] A. Tsanas, M. Little, P. McSharry, L. Ramig, Accurate telemonitoring of Parkinson's disease progression by non-invasive speech tests, Nature Precedings (2009) 1–1Publisher: Nature Publishing Group. `doi:10.`

1038/npre.2009.3920.1.
URL https://www.nature.com/articles/npre.2009.3920.1

[41] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, J. Reis, Modeling wine preferences by data mining from physicochemical properties, Decision support systems 47 (4) (2009) 547–553.

[42] R. Kohavi, others, Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid., in: Kdd, Vol. 96, 1996, pp. 202–207.

[43] C. A. Ratanamahatana, D. Gunopulos, Scaling up the naive bayesian classifier: Using decision trees for feature selection (2002).

[44] G. d'Aloisio, C. D. Sipio, A. D. Marco, D. D. Ruscio, Towards early detection of algorithmic bias from dataset's bias symptoms: An empirical study, Information and Software Technology 188 (2025) 107905. doi:https://doi.org/10.1016/j.infsof.2025.107905.
URL https://www.sciencedirect.com/science/article/pii/S0950584925002447

[45] A. Fabris, S. Messina, G. Silvello, G. A. Susto, Algorithmic fairness datasets: the story so far, Data Mining and Knowledge Discovery 36 (6) (2022) 2074–2152. doi:10.1007/s10618-022-00854-z.
URL https://doi.org/10.1007/s10618-022-00854-z

[46] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[47] Z. Chen, J. M. Zhang, F. Sarro, M. Harman, Fairness Improvement with Multiple Protected Attributes: How Far Are We?, conference Name: 2024 IEEE/ACM 46th International Conference on Software Engineering (ICSE) Meeting Name: 2024 IEEE/ACM 46th International Conference on Software Engineering (ICSE) Place: Lisbon, Portugal Publisher: IEEE/ACM Volume: 46 (Apr. 2024).
URL https://www.computer.org/csdl/proceedings/icse/2024/1RLIVDkr2xO

[48] G. Rosenfield, K. Fitzpatrick-Lins, A coefficient of agreement as a measure of thematic classification accuracy., Photogrammetric Engineering and Remote Sensing 52 (2) (1986) 223–227.
URL http://pubs.er.usgs.gov/publication/70014667

[49] M. Buckland, F. Gey, The relationship between recall and precision, Journal of the American society for information science 45 (1) (1994) 12–19, publisher: Wiley Online Library.

[50] R. F. Woolson, Wilcoxon signed-rank test, Encyclopedia of Biostatistics 8, publisher: Wiley Online Library (2005).

[51] F. Sarro, A. Petrozziello, M. Harman, Multi-objective software effort estimation, in: Proceedings of the 38th International Conference on Software Engineering, ICSE '16, Association for Computing Machinery, New York, NY, USA, 2016, pp. 619–630. doi:10.1145/2884781.2884830.
URL https://dl.acm.org/doi/10.1145/2884781.2884830

[52] A. Arcuri, L. Briand, A practical guide for using statistical tests to assess randomized algorithms in software engineering, in: Proceedings of the 33rd International Conference on Software Engineering, ICSE '11, Association for Computing Machinery, New York, NY, USA, 2011, pp. 1–10. doi:10.1145/1985793.1985795.
URL https://dl.acm.org/doi/10.1145/1985793.1985795

[53] S. Dutta, D. Wei, H. Yueksel, P.-Y. Chen, S. Liu, K. Varshney, Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing, in: International conference on machine learning, PMLR, 2020, pp. 2803–2813.