FACEGroup: Feasible and Actionable Counterfactual Explanations for Group Fairness

Christos Fragkathoulas[1,2](✉), Vasiliki Papanikou[1,2], Evaggelia Pitoura[1,2], and Evimaria Terzi[2,3]

[1] University of Ioannina pitoura@uoi.gr
[2] Archimedes, Athena Research Center, Greece {ch.fragkathoulas, v.papanikou}@athenarc.gr
[3] Boston University, USA evimaria@bu.edu

Abstract. Counterfactual explanations assess unfairness by revealing how inputs must change to achieve a desired outcome. This paper introduces the first graph-based framework for generating group counterfactual explanations to audit group fairness, a key aspect of trustworthy machine learning. Our framework, FACEGroup (Feasible and Actionable Counterfactual Explanations for Group Fairness), models real-world feasibility constraints, identifies subgroups with similar counterfactuals, and captures key trade-offs in counterfactual generation, distinguishing it from existing methods. To evaluate fairness, we introduce novel metrics for both group and subgroup level analysis that explicitly account for these trade-offs. Experiments on benchmark datasets show that FACEGroup effectively generates feasible group counterfactuals while accounting for trade-offs, and that our metrics capture and quantify fairness disparities.

Keywords: explanations · fairness · XAI · counterfactuals.

1 Introduction

AI-driven technologies increasingly shape critical decisions, making it essential to understand their underlying reasoning and evaluate their fairness. A variety of explanation methods have been proposed to enhance transparency [9,1], with counterfactual explanations (CFs) gaining prominence [34]. Individual CFs reveal how modifying specific features can alter model decisions, offering actionable insights. For example, consider a person whose loan application is rejected by a machine learning model; a CF might indicate that increasing annual income or reducing the debt-to-income ratio would lead to approval.

Prior work has primarily focused on individual counterfactual explanations (CFs) [18,16,10,33,27,29,25,31,4,2], with comparatively few studies addressing counterfactuals for groups of instances [28,23,19,20]. Group counterfactual explanations (GCFs) identify how a group of instances, often defined by shared characteristics or *protected attributes* such as sex or race, could collectively alter their features to achieve favorable outcomes. GCFs are not simply aggregations of individual CFs; rather, they reveal common patterns or barriers affecting the group as a whole, which is critical for understanding systemic disparities and informing policy or organizational decisions. Previous studies introduce group-based approaches, by identifying common patterns among individuals with favorable outcomes [28], learning global translation vectors, and scaling them for GCFs [23], or constructing decision trees via stochastic local search [19]. In contrast, our work is the first to generate GCFs using a graph-based approach that enforces feasibility, supports subgroup-level analysis, and explicitly addresses the key trade-offs involved in counterfactual generation.

FACEGroup, our approach for generating Feasible and Actionable Group Counterfactual Explanations (GCFs), generates GCFs using a density-weighted feasibility graph [27], where nodes represent data points and edges denote feasible transitions that comply with real-world constraints. To ensure plausibility, we restrict connections to allow only small feature changes between data points. A key property of this graph is that feasibility constraints, cost limitations, and density weighting naturally partition the data into weakly connected components (WCCs), effectively dividing each group into subgroups with similar feasible counterfactual explanations.

The generation of group counterfactual explanations (GCFs) inherently involves balancing several key trade-offs: the proportion of factual instances within a group that are explained by the selected set of counterfactuals (coverage), the effort or change required for group members to achieve a counterfactual (cost), and the number of unique counterfactuals generated for the group (interpretability). To address these trade-offs, we introduce two algorithmic formulations based on the feasibility graph: the cost-constrained approach, which maximizes group coverage under a cost limitation, and the coverage-constrained approach, which minimizes the maximum cost required to achieve a specified coverage level. Both formulations are supported by mixed-integer programming solutions and greedy heuristics that operate at both the group and subgroup levels. Our approach also ensures that the generated counterfactuals remain feasible and actionable.

Finally, we introduce novel fairness metrics for group counterfactuals, which enhance existing fairness measures by capturing the various trade-offs in counterfactual generation and can be applied at both group and subgroup levels. We evaluate FACEGroup on real-world datasets, showing its effectiveness in fairness auditing. Compared to existing methods, FACEGroup produces more feasible and compact counterfactuals that align with the data distribution.

The rest of this paper is structured as follows: Section 2 formalizes the problem, Section 3 presents our algorithms, Section 4 introduces our fairness measures, Section 5 details experiments, Section 6 discusses related work, and Section 7 concludes.

2 Problem Definition

Let $f : \mathbb{R}^d \rightarrow \{0, 1\}$ be a binary classifier which maps instances in a d-dimensional feature space into two classes, labeled 0 and 1. Let $U \subseteq \mathbb{R}^d$ denote the input space. A model prediction on an individual instance $\mathbf{x} \in U$, called *factual*, is explained by crafting a counterfactual (CF) instance $\mathbf{x}' \in \mathbb{R}^d$ that is similar to $\mathbf{x}$ but leads to a different outcome, i.e., $f(\mathbf{x}') \neq f(\mathbf{x})$ [34]. The changes in feature values from $\mathbf{x}$ to $\mathbf{x}'$ should be feasible and comply with real-world constraints, for instance, changes to immutable features, such as race or height, should be prohibited. Formally, a counterfactual $\mathbf{x}'$ for $\mathbf{x}$ is defined as: $\mathbf{x}' = \arg\min_{\mathbf{x}'' \in \mathcal{A}_\mathbf{x}} cost(\mathbf{x}, \mathbf{x}'')$ s.t. $f(\mathbf{x}'') \neq f(\mathbf{x})$, where $cost(\mathbf{x}, \mathbf{x}'')$ is a function measuring the cost of transitioning from $\mathbf{x}$ to $\mathbf{x}'$. The *feasibility set* $\mathcal{A}_\mathbf{x}$ denotes the set of counterfactuals attainable from $\mathbf{x}$ via feasible changes.

It would be hard to trust a CF if it resulted in a combination of features that were unlike any observations the classifier has encountered before [34]. Therefore, CFs should also be coherent with the underlying data distribution. To ensure both feasibility and plausibility, we adopt a graph-based approach. Following [27], we construct a weighted directed graph $G_U = (V, E, W)$. Nodes correspond to instances in U, and an edge from node $\mathbf{x_i}$ to node $\mathbf{x_j}$ represents a feasible transition in the feature space. We call this graph *feasibility graph*. Transitions are further constrained by a cost threshold ϵ, ensuring that only small-cost feature changes are allowed. This ensures that changes between instances are both feasible and small. The weight function W is defined using a density-based approach [27] to ensure that CFs lie in dense areas of the input space and avoid outliers. Each edge in G_U is assigned a weight W_{ij}, calculated as the product of the density of the instances around the midpoint of $\mathbf{x_i}$ and $\mathbf{x_j}$ estimated using a Kernel Density Estimator (KDE) [8], and the cost between instances: $W_{ij} = KDE\left(\frac{\mathbf{x_i}+\mathbf{x_j}}{2}\right) cost(\mathbf{x_i}, \mathbf{x_j})$.

Given G_U, we now formally define the feasibility set $\mathcal{A}_\mathbf{x}$ of factual $\mathbf{x}$ as the set of instances $\mathbf{x}'$ for which there is a path in G_U from $\mathbf{x}$ to $\mathbf{x}'$, i.e., the set of instances that are reachable from $\mathbf{x}$: $\mathcal{A}_\mathbf{x} = \{\mathbf{x}' \in \mathbf{U} | \mathbf{x}' \text{ is reachable from } \mathbf{x} \text{ in } G_U\}$. These instances are the *feasible* CFs for $\mathbf{x}$.

Instead of finding a CF for a single factual $\mathbf{x}$, we are interested in providing CFs for a set $X \subseteq U$ of instances mapped to the same class. Let $X' \subseteq U$ be the set of instances mapped to the opposite class. Our goal is to identify a small subset S of X' of size k that best explains X. We limit the number of CFs to k for interpretability. To select S, we consider coverage-cost trade-offs. For a set of CFs $S \subseteq X'$, coverage is:

$$coverage(X, S) = |\{\mathbf{x} \,|\, \mathbf{x} \in X \text{ and } \exists\, \mathbf{x}' \in S \cap \mathcal{A}_\mathbf{x}\}| \,.$$

We overload the notation for *cost* to define the cost between an instance and a set, as well as between two sets:

$$cost(\mathbf{x}, S) = \min_{\mathbf{x}' \in S} cost(\mathbf{x}, \mathbf{x}'), \quad cost(X, S) = \min \max_{\mathbf{x} \in X} cost(\mathbf{x}, S).$$

The function $cost(\mathbf{x}, \mathbf{x}')$ captures the cost of transforming $\mathbf{x}$ to $\mathbf{x}'$, offering flexibility to adapt to specific problem requirements. For example, cost can be defined as the vector distance (e.g., L2 norm), the sum of edge weights along the shortest path in G_U, or simply the number of hops on this path. By emphasizing proximity in feature space and by considering dense paths, these definitions ensure that the CFs are closely aligned with the data distribution. Our approach works with any definition of cost.

A necessary condition for $\mathbf{x}'$ to be a feasible counterfactual for $\mathbf{x}$ is that both $\mathbf{x}$ and $\mathbf{x}'$ belong to the same weakly connected component (WCC) of G_U. As a result, G_U induces a partition of the set of factual instances X into m disjoint subsets $X_1, \ldots, X_m$, $m > 0$. Each subset X_i contains instances in X that belong to the same WCC of G_U and thus share a common space of feasible counterfactuals, denoted X'_i, which also reside within the same component. This partitioning of X into subgroups with distinct feasible counterfactual spaces offers a meaningful perspective for analyzing model behavior at both the group and subgroup level, highlighting regions of the input space that support similar feasible explanations.

We now provide two definitions of the FACEGroup problem. Our first definition prioritizes cost over coverage, setting a threshold on cost, and our second definition prioritizes coverage over cost, asking for a set that provides a specified coverage degree c.

Problem 1 (Cost-Constrained). Given X, X', $k \in \mathbb{N}^*$, and cost threshold $d \in \mathbb{R}^+_*$, find $S \subseteq X'$ with $|S| \leq k$ and $Q \subseteq X$ such that for every instance $\mathbf{x} \in Q$ there exist an instance $\mathbf{x}' \in S$ such that $cost(\mathbf{x}, \mathbf{x}') \leq d$ and $|Q|$ is maximized.

Problem 2 (Coverage-Constrained). Given X, X', $k \in \mathbb{N}^*$, and coverage degree c, $0 < c \leq 1$, find $S \subseteq X'$ with $|S| \leq k$ such that $coverage(X, S) \geq c\,|X|$ and $cost(X, S)$ is minimized.

3 Algorithms

Our approach to generating feasible CFs is based on the feasibility graph G_U. Both optimization problems are NP-hard. The cost-constrained problem can be formulated as an instance of the maximum coverage problem, while the coverage-constrained problem is similar to the classical k-center problem [32].

In the following, we present two versions for both problems: (a) a global version that generates CFs for the whole set X and (b) a local version that generates CFs per subgroup X_i. We also show how the local version can be used to generate CFs for the whole group X. A common step in both problems involves computing, for each factual $\mathbf{x}$, the candidate counterfactuals, i.e., the feasibility set $\mathcal{A}_{\mathbf{x}}$ and computing costs. To this end, we use Breadth-First-Search for vector costs (e.g., L2 distance) and Dijkstra's algorithm for shortest path costs, with complexities of $O(|V| + |E|)$ and $O(|V| \log |V| + |E|)$, respectively.

3.1 The Cost-Constrained FACEGroup Problem

We solve this problem using two approaches: (a) a Mixed-Integer Programming (MIP) that explicitly models constraints for each factual-counterfactual pair while optimizing coverage, and (b) a Greedy approach that iteratively selects CFs to maximize coverage.

For the MIP solution of the global version of the problem, we define two binary decision variables. Let $r_{\mathbf{x}\mathbf{x}'} = 1$ if $\mathbf{x}'$ covers $\mathbf{x}$; and $r_{\mathbf{x}\mathbf{x}'} = 0$, otherwise, and $u_{\mathbf{x}'} = 1$ if CF $\mathbf{x}'$ covers any instance in X, and $u_{\mathbf{x}'} = 0$ otherwise. The goal is to maximize the number of covered factual instances:

$$\max \sum_{\mathbf{x}' \in X'} \sum_{\mathbf{x} \in X} r_{\mathbf{x}\mathbf{x}'} \quad \text{s.t.} \quad \sum_{\mathbf{x}' \in X'} u_{\mathbf{x}'} \leq k \quad (1)$$

$$\sum_{\mathbf{x}' \in X'} r_{\mathbf{x}\mathbf{x}'} \leq 1, \ \forall \mathbf{x} \in X \quad (2) \quad r_{\mathbf{x}\mathbf{x}'} \leq u_{\mathbf{x}'}, \ \forall \mathbf{x}' \in X', \forall \mathbf{x} \in X \quad (3),$$

$$u_{\mathbf{x}}, \, r_{\mathbf{x}\mathbf{x}'} \in \{0, 1\}, \ \forall \mathbf{x} \in X, \ \mathbf{x}' \in X'. \quad (4)$$

While constraint (1) limits the number of selected CFs to at most k, constraint (2) enforces that each factual instance $\mathbf{x}$ is assigned to at most one CF $\mathbf{x}'$. Constraint (3) guarantees that if a CF $\mathbf{x}'$ is assigned to cover a factual instance $\mathbf{x}$ ($r_{\mathbf{x}\mathbf{x}'} = 1$) then $\mathbf{x}'$ must be selected $u_{\mathbf{x}'} = 1$, and constraint (4) defines the binary decision variables. This formulation has $O(2^{|X'|})$ complexity.

For the global Greedy version of the problem, we iteratively select counterfactuals (CFs) to maximize coverage. Let S_t be the set of counterfactuals selected at iteration t. We start with an empty set $S_0 = \emptyset$. At each iteration t, the algorithm selects the CF $\mathbf{x}' \in X'$ that

$$\mathbf{x}' = \arg \max_{\mathbf{x}'' \in X'} (coverage(X, S_{t-1}) + coverage(X, \{\mathbf{x}''\}), \quad (5)$$

updates $S_t = S_{t-1} \cup \{\mathbf{x}'\}$, and terminates when either $|S_t| = k$ or all instances in X are covered.

The worst-case complexity of this algorithm is $O(k|X|)$. Given the submodular nature of coverage, where the marginal gain of adding a new CF to the set S decreases as S grows, it adheres to the properties of submodular maximization. Consequently, the attained coverage is no worse than $(1 - \frac{1}{e})$ times the optimal maximum coverage [17].

The Greedy algorithm can also be used to provide a counterfactual explanation for a subgroup X_i by applying it only to the corresponding WCC. We can also utilize this local version to provide counterfactuals for the whole group X by applying the Greedy algorithm iteratively to all m WCC as follows. Initially, we apply a single step of the Greedy algorithm at each WCC. Then, we select the CF that provides the best coverage and apply an additional step of the algorithm to the WCC from which the CF was selected. We repeat this until the maximum number k of counterfactuals is reached or all factual instances are covered. It is easy to see that this local version provides the same result as the global one. The local Greedy selection has the same complexity as the global Greedy approach, as it follows a similar process while iterating over WCCs, either scanning all $|X'|$ candidates or evaluating coverage within each component.

3.2 The Coverage-Constrained FACEGroup Problem

To solve this problem, we employ two algorithms: a mixed-integer programming (MIP) and a Greedy 2-approximation algorithm [13]. While the Greedy algorithm provides an efficient yet approximate solution, the MIP guarantees optimal results [7], but can become computationally expensive for large graphs.

For the MIP formulation, the solution is similar to the Cost-Constrained problem with the following modifications. The objective function minimizes the maximum cost d of the farthest instance while ensuring that $coverage(X, S) \geq c\,|X|$. Constraints (1), (2), (3), and (4) still apply, along with:

$$\sum_{\mathbf{x}' \in X'} cost(\mathbf{x}, \mathbf{x}') r_{\mathbf{x}\mathbf{x}'} \leq d, \quad \forall \mathbf{x} \in X \quad (6), \qquad \sum_{\mathbf{x}' \in X'} \sum_{\mathbf{x} \in X} r_{\mathbf{x}\mathbf{x}'} \geq c\,|X| \quad (7).$$

Constraint (6) ensures that the cost of any node to its assigned center does not exceed d, enforcing the objective function, and Constraint (7) enforces that the desired coverage percentage is achieved. For full coverage, $c = 1$, constraint (2) becomes an equality constraint, and constraint (7) is no longer needed.

For the Greedy algorithm, the process begins by arbitrarily selecting the first counterfactual $\mathbf{x}'$ and assigning all factuals $\mathbf{x}$ within a cost of r to it, where r is initially set to the maximum cost between any factual and candidate counterfactual. We then iteratively select the counterfactual that is farthest from those already chosen and assign all factuals within a cost of r to it. This process continues until we reach the predefined coverage or the number of counterfactuals k. To find the smallest value of r that satisfies the coverage requirement, we employ a binary search. The complexity of this algorithm is $O(k|X|log(d))$, since it assigns up to $|X|$ factuals for each of the k selected counterfactuals and binary search adds this logarithmic factor $log(d)$, where d is the range of costs considered.

Both the MIP and the Greedy approaches can be applied globally and locally. In the global version, we apply the algorithms on the G_U graph. In the local version, for a specific subgroup X_i of X, the algorithms are applied within the corresponding WCC of G_U.

We now describe how the local version can be used to solve the global version. Consider the case of full coverage ($c = 1$) with m WCCs ordered arbitrarily as $C_1, C_2, \ldots, C_m$. Achieving full coverage reduces to distributing k counterfactuals among these components. Since at least one counterfactual is required per WCC, the maximum allocation per WCC is at most $k - m$. First, we run MIP or Greedy within each WCC, varying k from 1 to $k - m$. Let l_i be the minimum counterfactuals needed to fully cover C_i. We start by assigning l_i to each C_i, then iteratively allocate remaining counterfactuals to the WCC with the highest cost until the total reaches k.

When $c < 1$, the task becomes more complex as we have to allocate both k and coverage c across the WCCs. Let $F(1...i, k, n)$ be the minimum cost of allocating k counterfactuals that cover a total of n factuals considering connected components $WCC_1, ..., WCC_i$, where $n = c|X|$. Similarly, let $F(i, k, n)$ represent the minimum cost of allocating k counterfactuals to cover n factuals within component WCC_i. Then, we can solve the problem with time complexity of $O(m(kn)^2)$, using dynamic programming as follows:

$$F(1...i, k, n) = \min_{1 \leq n' \leq n, 1 \leq k' \leq k} \{F(1...i-1, k-k', n-n') + F(i, k', n')\}$$

For large graphs, solving the MIP at a global level can become computationally demanding, as the number of decision variables and constraints grows exponentially with the dataset size. To improve performance, we add constraints only for instances $\mathbf{x}$ and $\mathbf{x}'$, such that $\mathbf{x}' \in A_X$, reducing unnecessary computations. For full coverage, the complexity of the global Greedy approach is $O(|X|k\log(d))$ while the complexity for the local approach is $O(m(k-m)|X_i|k\log(d))$.

4 FACEGroup for Auditing Fairness

In this section, we examine algorithmic fairness through the lens of FACEGroup. Group fairness refers to a set of principles designed to ensure that protected groups, often defined by sensitive attributes such as gender, race, or age, are treated similarly by a classifier. Broadly, group fairness can be categorized into *demographic parity*, which requires that the proportion of positive outcomes reflects representation of the group in the population, and *error-based fairness*, which focuses on equalizing classification errors, such as false negative rates, across groups [36,9].

To audit fairness for a group X, we generate group counterfactual explanations (GCFs) for relevant subsets of X. For example, we generate GCFs for the negatively classified instances of X when auditing for demographic parity, or the false negatives of X when auditing for error-based fairness. Disparities in the GCFs generated for different groups (e.g., males vs. females) can reveal potential biases in the model.

Unlike existing approaches, FACEGroup supports *multi-level* fairness auditing by partitioning each group into subgroups according to the connected components of the feasibility graph. This allows us to examine unfair behavior not only at the group level, but also at the level of subgroups, offering finer-grained insight into patterns of bias. Furthermore, to capture the *key trade-offs* in generating counterfactuals, FACEGroup provides novel fairness metrics that are parameterized by the number k of counterfactuals, the cost d, and the coverage c. Introducing the number k in the fairness metrics allows for assessing interpretability, as groups requiring fewer CFs are more interpretable, it promotes trust, as models that require fewer CFs are more transparent, and it serves in detecting disparities in CF requirements across (sub)groups, factors previously overlooked.

Burden-based Fairness Measures. Counterfactuals provide a novel approach to measuring unfairness by evaluating both the disparities in outcomes between groups and the effort required by these groups to achieve fairness, i.e., to obtain the positive outcome. This effort, also called *burden*, is often estimated as the aggregated cost between the factuals in a group and their counterfactuals

[31,22]. However, measuring burden solely at the group level may obscure disparities within subgroups, as different subpopulations may face varying degrees of difficulty in achieving favorable outcomes.

We first define the minimum k (k_0) and cost (d_0) required for full coverage ($c = 1$):

$$k_0 = \min\{k \mid \exists S, |S| \leq k, \text{coverage}(X, S) = |X|\},$$
$$d_0 = \min\{d \mid \exists S, \text{cost}(X, S) \leq d, \text{coverage}(X, S) = |X|\}.$$

Note that k_0 is lower-bounded by the number of weakly connected components ($k_0 \geq m$), and d_0 does not exceed the largest WCC diameter.

We now introduce *AUC-based fairness measures* that assess trade-offs between cost, number of counterfactuals, and coverage of (sub)groups across a range of parameter values rather at fixed points, avoiding biases from rigid parameter settings. The corresponding *saturation points* identify optimal thresholds for cost, number of counterfactuals, and coverage.

We define the set of counterfactuals $S_{k,d}$ that maximize coverage under a cost constraint d as:

$$S_{k,d} = argmax_{|S| \leq k,\, cost(X,S) \leq d} |coverage(X, S)|$$

and $kAUC(k)$ as:

$$kAUC(k) = \int_{d_{min}}^{d_{max}} \text{coverage}(X, S_{k,d})\, dd$$

that measures how efficiently a group can achieve coverage across a range of cost values for a given number of counterfactuals.

Similarly, we define $dAUC(d)$ to evaluate how coverage improves as the number of counterfactuals increases under a fixed cost constraint, and $cAUC(c)$ to quantify the effort required to reach a given coverage level by measuring the total cost over a range of counterfactual numbers. Figure 1 provides a visual representation of the AUC-based metrics.

There is also a minimum cost that provides the highest attainable coverage for k, we call it *saturation point* for k and denote it as $sp(k)$. Formally, it holds, for any $d \geq sp(k)$, $coverage(X, S_{k,d}) = coverage(X, S_{k,sp(k)})$. Similarly, we define, $sp(d)$ to determine the least number of counterfactuals needed to reach maximum coverage within a given cost constraint, and $sp(c)$ to represent the minimum cost needed to achieve a desired coverage level, helping quantify the burden on different groups. Saturation points are shown in Figure 1.

Attribution Measures. FACEGroup also provides insights into feature importance by measuring how often a feature change is required to alter an outcome. Concretely, the *attribute change frequency* (ACF) metric captures how frequently a feature A changes between a factual instance $\mathbf{x} \in X$ and its corresponding counterfactual $\mathbf{x}' \in S$:

$$ACF(X, S, A) = \frac{1}{|X|} \sum_{\mathbf{x} \in X} (1 - \delta(\mathbf{x}_A, \mathbf{x}'_A)),$$

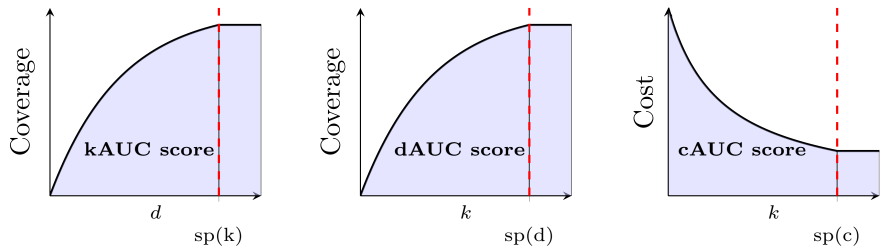


Fig. 1: AUC scores and saturation points

where $\delta(\mathbf{x}_A, \mathbf{x}'_A)$ is the Kronecker delta, returning 1 if the feature remains unchanged and 0 otherwise. and $\mathbf{x}_A$ and $\mathbf{x}'_A$ represent the values of A in the factual and counterfactual instances, respectively. For each factual instance, we get the corresponding counterfactual instance with the minimum cost, i.e., $\mathbf{x}' = argmin_{\mathbf{x}'' \in S} cost(\mathbf{x}, \mathbf{x}'')$.

5 Experimental Evaluation

The goal of our experimental evaluation is twofold: (a) to demonstrate the effectiveness of FACEGroup in fairness auditing and (b) to compare FACEGroup with baseline group counterfactual methods.

For fairness auditing, we use the widely studied `Adult`[4] dataset for income classification. To benchmark FACEGroup with baselines, we extend evaluations to additional datasets derived from US Census surveys, `AdultCA`[5], `AdultLA`[2], and other domains including `COMPAS`[6], `Student`[7], `German Credit`[8], and `HELOC`[9]. Further details on preprocessing, parameter settings, and configurations, as well as additional experiments on other datasets, are in the supplementary material. The source code is available online[10].

First, we construct the feasibility graph G_U. An edge exists from a $\mathbf{x_i}$ to a $\mathbf{x_j}$ if the transition from $\mathbf{x_i}$ to $\mathbf{x_j}$ is feasible and within threshold ϵ. We use a small set of generic feasibility constraints prohibiting unrealistic modifications, such as changing the values of immutable attributes (e.g., race) or the directionality of others, such as decreasing the value of the age attribute. The full set of constraints used is in the supplementary material. We define groups based on the sensitive attribute *Gender*: G_0 (females) and G_1 (males).

Figure 2 depicts the impact of varying ϵ on graph connectivity metrics, showing values up to the point where nearly all instances are connected, minimizing singleton nodes. Smaller ϵ values result in sparser graphs, ensuring that connected instances are more similar, leading to more plausible, small-step transitions. Conversely, larger ϵ values create denser graphs by incorporating connections between more distant instances, allowing for larger transition steps. To

[4] Adult [5] Adult-CA-LA Datasets [6] COMPAS [7] Student [8] German Credit
[9] HELOC [10] Project Repository

balance plausibility with connectivity, we select the smallest possible ϵ that maintains a highly connected graph while minimizing singleton nodes. For the `Adult` dataset, we set $\epsilon = 0.4$. Further results for the selection of ϵ on the remaining datasets can be found in the supplementary material.

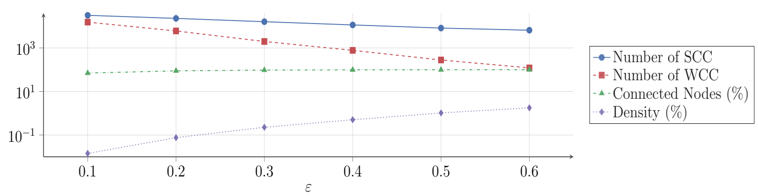


Fig. 2: Feasibility graph connectivity based on the ϵ constraint.

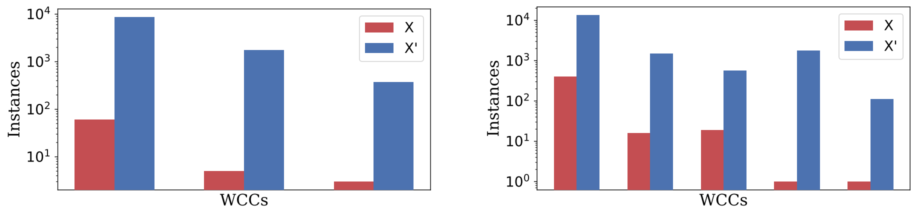
5.1 Auditing Fairness

In this set of experiments, we apply our algorithms to audit fairness. Without loss of generality, we focus on finding GCFs for the negatives for both groups G_0 and G_1. We use an XGBoost classifier optimized via hyperparameter tuning. We consider only the instances in G_0 and G_1 for which at least one feasible candidate CF exists and use the L_2 distance as the cost function.

Burden Analysis. A key strength of FACEGroup is its ability to uncover subgroup behaviors within the groups G_0 and G_1 through the feasibility graph G_U, which naturally partitions each group into WCCs, representing subpopulations that share feasible CF transformations. Figure 3 visualizes the distribution of factual instances (X, red) and feasible counterfactual candidates (X', blue) across the subgroups (WCCs) of each group. We observe that G_1 exhibits a more fragmented structure, with CFs more widely spread across subgroups compared to G_0, suggesting that G_1 has a higher degree of variability in the transformations required for favorable outcomes. Table 1 depicts the minimum resources (k_0 and d_0) needed for full coverage per subgroup (WCC). G_1 requires more CFs ($k_0 = 12$) than G_0 ($k_0 = 9$) and higher minimum cost ($d_0 = 1.04$) than G_0 ($d_0 = 0.93$), suggesting greater heterogeneity in the CF pathways needed for full coverage.

Analyzing subgroups is crucial, as group-level fairness assessments can mask heavily disadvantaged subpopulations, leading to misleading conclusions about the equitable distribution of the burden. At the subgroup level, the *Black* subgroups (that correspond to WCC_1 in both groups) exhibit the highest k_0 and d_0, indicating that they face greater barriers to obtain favorable decisions. Notably, the subgroups with the most factual instances also bear the highest burden, indicating a disproportionate impact on overall group difficulty.

Table 2 reports $kAUC$, $dAUC$, $cAUC$, saturation points sp, and the minimum, or maximum values for coverage and cost, that correspond to each sp.



(a) G_0 (b) G_1

Fig. 3: Distribution of $\mathbf{X}$, $\mathbf{X}'$ per WCC of the subgroups G_0 and G_1.

Table 1: k_0 and d_0 for each WCC of each group and overall for each group.

	WCCs										Overall	
	WCC_1		WCC_2		WCC_3		WCC_4		WCC_5		k_0	d_0
	k_0	d_0	k_0	d_0	k_0	d_0	k_0	d_0	k_0	d_0		
G_0	7	0.93	1	0.74	1	0.49	–	–	–	–	9	0.93
G_1	4	1.04	3	0.61	3	0.78	1	0.46	1	0.20	12	1.04

Scores are normalized by the optimal AUC per metric. Higher $kAUC$, $dAUC$ and lower $cAUC$ are preferred.

For $kAUC$, saturation points (sp) are expected to decrease as more CFs are provided. Initially, at $k = 1$, G_1 achieves higher maximum coverage, reflecting larger available transitioning costs, enabling more instances to be efficiently covered at low k. However, as the number of CFs increases, G_0 reaches full coverage first, exhibiting better overall efficiency (higher $kAUC$) and requiring fewer resources (lower sp values) compared to G_1. For $dAUC$, saturation points should decrease as higher-cost connections are allowed. At $d = 0.1$, G_0 has a lower $sp(d)$, indicating fewer feasible low-cost available transitions, compared to G_1. As cost increases, G_0 effectively utilizes connections to reach full coverage with fewer CFs, while G_1 requires higher costs to achieve maximum comparable coverage. However, when $d \in [0.8, 1.5]$, G_1 exhibits stronger coverage efficiency gains, suggesting G_0 is more efficient at lower costs while G_1 benefits more from cost relaxations. For $cAUC$, both groups experience similar cost burdens for achieving intermediate coverage levels $0.25, 0.5$ and 0.75. However, at full coverage ($c = 1.0$), G_1 incurs significantly higher costs, as reflected in both $cAUC$ and minimum cost. The consistently higher $sp(c)$ values for G_1 suggest that more CFs are required to reach cost-efficient solutions, reinforcing a systemic disadvantage in obtaining full coverage at minimal cost while maintaining interpretability.

Attribution Analysis. To further analyze subgroup disparities, we use the ACF metric per WCC, quantifying how often specific features are altered in CFs, providing insights into the different factors driving classification decisions. Figure 4 presents the frequency of modified attributes for each WCC of G_0 and G_1,

Table 2: $kAUC$, $dAUC$, $cAUC$, and saturation points.

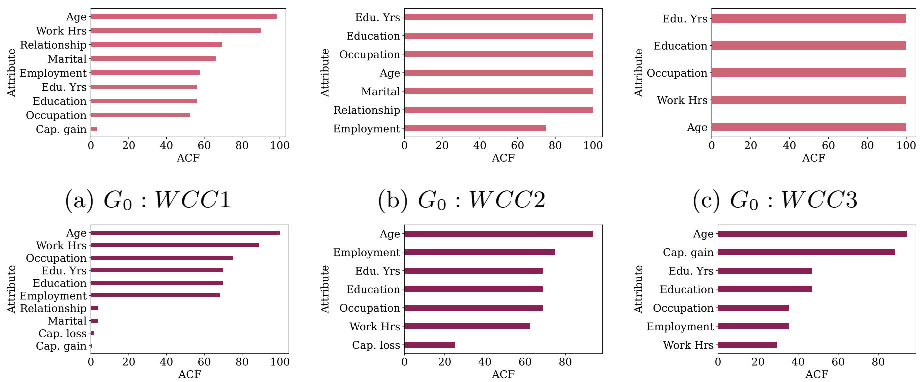
Parameter	**Value**	G_0			G_1		
				kAUC metrics			
		$sp(k)$	Max Cov.	$kAUC$	$sp(k)$	Max Cov.	$kAUC$
k	1	1.1	63.08	0.50	1.3	65.75	0.54
	5	1.1	93.85	0.82	1.1	97.49	0.85
	9	1.1	100.0	0.90	1.1	99.09	0.89
	13	0.7	100.0	0.92	1.1	100.0	0.91
				dAUC metrics			
		$sp(d)$	Max Cov.	$dAUC$	$sp(d)$	Max Cov.	$dAUC$
d	0.1	6	12.31	0.10	12	12.78	0.08
	0.8	10	100.0	0.89	12	99.31	0.93
	1.5	9	100.0	0.93	12	99.77	0.95
	2.2	9	100.0	0.93	12	99.77	0.95
				cAUC metrics			
		$sp(c)$	Min Cost	$cAUC$	$sp(c)$	Min Cost	$cAUC$
c	0.25	12	0.14	0.10	20	0.12	0.11
	0.50	18	0.22	0.17	23	0.20	0.17
	0.75	22	0.28	0.25	25	0.30	0.25
	1.00	16	0.55	0.56	20	1.40	0.72

respectively, and shows that subgroup-specific variations exist in the importance of different features. For G_1, we include only the three largest WCCs, excluding those with few factual instances, as they lack representativeness. A common trend across all WCCs in both groups is that an increase in *age* is frequently required for a favorable outcome, suggesting that the model associates age with work experience or financial stability. Within G_0, the Asian-Pacific-Islander individuals (WCC_3) require fewer modifications compared to the Blacks (WCC_1) and Amer-Indian-Eskimos (WCC_2) and do not rely on *relationship status* or *marital status*, unlike the others. In G_1, despite similar CF difficulty (Table 1), financial interventions differ: Amer-Indian-Eskimos (WCC_2) require career-related changes (*employment status, occupation, education*), while Asian-Pac-Islanders (WCC_3) depend on increasing *capital gain.* More broadly, *capital gain* is largely absent from both groups of CFs except for $G_1 - WCC_3$, highlighting subgroup differences in financials to favorable outcomes. Finally, CFs in G_1 rarely modify *relationship status*, unlike in G_0, where it is frequently altered. Instead, *educational* and *occupational* factors are highly important.

5.2 Comparison with Baselines

We evaluate FACEGroup against existing CF generation methods, specifically: (a) with FACE [27], a graph-based method for individual CFs, and (b) with AReS [28] and GLOBE-CE [23], two state-of-the-art GCF approaches.

Comparison with Individual CFs Given a group X, FACEGroup generates a small set S of k counterfactuals to cover X. To evaluate the efficiency



(a) $G_0 : WCC1$ (b) $G_0 : WCC2$ (c) $G_0 : WCC3$

(d) $G_1 : WCC1$ (e) $G_1 : WCC2$ (f) $G_1 : WCC3$

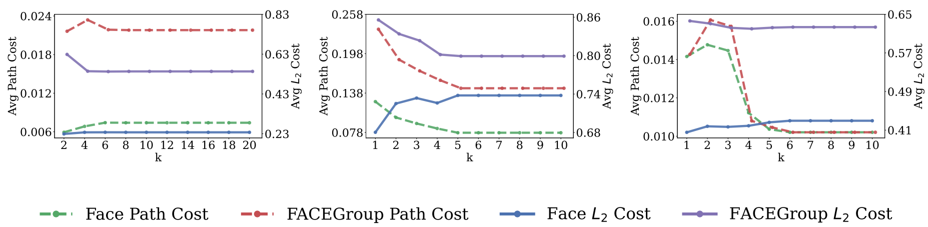
Fig. 4: ACF across the subgroups of each group

of this approach, we compare the associated cost with the cost of generating *individual counterfactuals* for each instance in X, which serves as a lower bound on the cost when the constraint on k is relaxed. For generating individual counterfactuals, we use FACE, since it is also based on a feasibility graph. For these experiments, we generate CFs for the full population $G = G_0 \cup G_1$. We assess how closely GCFs from FACEGroup approximate the optimal costs of individual CFs from FACE. First, we apply FACEGroup to generate the set S of CFs by solving the coverage-constrained problem. Then, we apply FACE to all factuals covered by S using the same cost function. As a cost function, we use both: (a) the weighted shortest path cost in G_U (originally used in FACE), and (b) the L_2 distance.

Figure 5 shows the cost comparison for k CFs from 1 to k_0 in 10 equal steps, with normalized costs. As expected, FACE achieves the lowest costs, while FACEGroup, which prioritizes group-level explanations, incurs slightly higher but still near-optimal costs. FACEGroup maintains near-optimal shortest path costs in datasets like `German Credit` and `HELOC`, where feasible transformations remain efficient. However, in `Adult`, costs increase due to the challenge of balancing feasibility with compact group CFs. Similar trends hold across other datasets, with full results and parameter details provided in the supplementary material.

Comparison with GCF Methods We compare FACEGroup with two state-of-the-art GCF baselines: AReS [28] and GLOBE-CE [23]. AReS mines frequent itemsets from individuals who achieved the desired outcome, selecting a small, interpretable set of rules via a submodular objective. GLOBE-CE defines global CFs as translation vectors applied to groups, scaling them across a range of values to adapt to individuals.

Both baselines without feasibility and plausibility constraints achieve at least 70% coverage. AReS generates 3 to 20 rules, while GLOBE-CE produces a significantly larger set, ranging from 10 to 612 CFs, due to the multiple scales on



(a) **Adult** Dataset (b) **German Credit** Dataset (c) **HELOC** Dataset

Fig. 5: Comparison of FACEGroup and FACE on average CF costs.

Table 3: Comparison with baselines.

Dataset	ϵ	FC	AReS		GLOBE-CE		FACEGroup	
			r	Cov. (%)	k	Cov. (%)	k	Cov. (%)
Adult	0.4	all	18	15.68	421	0.24	21	100
	0.4	none	18	52.26	421	84.56	10	100
AdultCA	0.7	all	20	11.36	612	11	133	100
	0.7	none	20	11.36	612	11.50	15	100
AdultLA	0.5	all	20	12.9	342	12.9	59	100
	0.5	none	20	23.11	342	22.63	13	100
Student	3.0	all	3	33.3	10	50	3	100
	3.0	none	3	75	10	66.67	2	100
COMPAS	0.3	all	20	11.85	124	20	13	100
	0.3	none	20	16.3	124	25.93	13	100
German Credit	2.9	all	4	0	18	26.32	6	100
	2.9	none	4	42.11	18	73.68	2	100
HELOC	1.4	all	11	1.98	74	1	4	100
	1.4	none	11	71.29	74	72.28	2	100

top of the translation vectors. Detailed results are in the supplementary material. To assess feasibility, we integrate CFs into the feasibility graph G_U and measure feasibility coverage as the proportion of CFs with at least one feasible transition. We analyze this under all feasibility constraints and a relaxed setting with only the plausibility constraint ϵ. Table 3 highlights the limitations of baselines: with full constraints, AReS and GLOBE-CE remain below 50% feasibility coverage, indicating that many CFs violate real-world constraints. In contrast, FACEGroup achieves 100% feasibility coverage with a compact CF set. Relaxing constraints improves coverage for baselines, particularly for GLOBE-CE, which benefits from its low-cost translation vectors. However, FACEGroup still maintains full feasibility coverage with fewer CFs, demonstrating its ability to generate feasible, actionable CFs without sacrificing interpretability or plausibility.

6 Related Work

Explanations have become central in machine learning research [9,14], particularly in high-stakes domains such as healthcare and education. Among various explanation methods, CFs have gained prominence for their ability to reveal actionable changes leading to a desired outcome. Wachter et al. [35] first formulated CFs as an optimization problem, minimizing the cost between an instance and its CF while ensuring a prediction change. Subsequent work [15,25,18,12,34,31,27] refined CF generation, emphasizing properties such as feasibility, actionability, sparsity [34], and robustness[16]. Several approaches optimize CF search using genetic algorithms [31,10], integer programming [30,33], and cost-based heuristics [12].

FACE [27] constructs a density-weighted feasibility graph where counterfactuals are generated via shortest paths in the graph, focusing on individual explanations that balance proximity and data manifold alignment. While FACEGroup builds on this graph structure, and further introduces three key innovations: (1) multi-level subgroup analysis, where WCCs of the feasibility graph naturally partition groups into interpretable subgroups with shared feasibility constraints, (2) GCF trade-off-aware algorithms, rather than relying on individual shortest-path searches, and (3) cost function agnosticism.

While most methods focus on individual CFs, recent work explores GCFs for multiple instances. AReS [28] defines subgroup-specific CF rules, optimizing for correctness, coverage, cost, and interpretability. GLOBE-CE [23] learns global translation vectors, applying them at different scales to generate CFs that maximize coverage. CET [19] uses decision trees for group actions to enhance transparency and consistency, while mixed-integer programming has been used to optimize collective CFs under linking constraints [5]. CounterFair [21] generates fair GCFs by selecting a subset via mixed-integer programming to balance cost and fairness. Unlike these approaches, FACEGroup enforces feasibility constraints, ensuring GCFs adhere to real-world constraints. Most group-based methods only prevent changes in sensitive attributes but lack directional constraints, leading to CFs that may violate plausible transformations. Notably, GLOBE-CE selects random feature perturbations, which can result in unrealistic CFs. In contrast to these methods, FACEGroup generates CFs at both group and subgroup levels, systematically handling the trade-offs in CF generation.

Explanations are utilized to assess algorithmic fairness [9], ensuring decisions are not influenced by protected attributes [26,24,11,6]. Several CF-based approaches have been proposed to quantify fairness by measuring the burden quantified as the difficulty individuals face in achieving a favorable outcome per group [22,31,12,20,28]. Methods like [31,22] generate individual CFs and calculate burden per group as the average sum of pairwise costs to assess fairness. PreCoF [12] distinguishes between explicit bias, when individual counterfactuals require changes only in sensitive attributes, and implicit bias, when, after removing sensitive attributes from model training, other features disproportionately influence different groups. [28,23] suggest that generated rules and global translation vectors can be used to manually audit for unfairness in subgroups of interest. FACTS [20] builds on AReS and introduces burden-based fairness metrics, but evaluates fairness only under specific settings. For instance, its Equal Cost of Effectiveness metric compares the minimum cost needed for protected subgroups to reach a fixed aggregate effectiveness level, defined as the proportion of individuals able to achieve the desired outcome via counterfactuals. In contrast, our burden-based fairness metrics assess disparities across a range of costs, coverage levels, and numbers of counterfactuals, offering a more comprehensive perspective that captures potential disparities across various combinations of these factors. Unlike the other approaches, FACEGroup introduces fairness metrics that assess fairness at both group and subgroup levels, explicitly accounting for trade-offs between cost, coverage, interpretability, and feasibility.

7 Conclusions

In this paper, we propose FACEGroup, a novel graph-based framework for group counterfactual generation that addresses limitations in existing methods by incorporating real-world feasibility constraints and managing trade-offs in counterfactual generation. We also introduce novel fairness measures that allow auditing fairness both at the group and subgroup levels, offering insights on the trade-offs between cost, the number of generated counterfactuals, and coverage. In future work, we plan to extend the use of the feasibility graph to define path-based fairness metrics. We also aim to adapt our approach to multi-class classification and regression settings.

8 Acknowledgment

This work has been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program.

References

1. Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., et al.: Explainable ai (xai): Core ideas, techniques, and solutions. ACM Computing Surveys **55**, 1–33 (2023)
2. Karimi, A.H., Barthe, G., Balle, B., Valera, I.: Model-agnostic counterfactual explanations for consequential decisions. In: International conference on artificial intelligence and statistics. pp. 895–905. PMLR (2020)
3. Slack, D., Hilgard, A., Lakkaraju, H., Singh, S.: Counterfactual explanations can be manipulated. Advances in neural information processing systems **34**, 62–75 (2021)
4. Bynum, L.E., Loftus, J.R., Stoyanovich, J.: Counterfactuals for the future. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 14144–14152 (2023)
5. Carrizosa, E., Ramírez-Ayerbe, J., Morales, D.R.: Generating collective counterfactual explanations in score-based classification via mathematical optimization. Expert Systems with Applications **238**, 121954 (2024)

6. Caton, S., Haas, C.: Fairness in machine learning: A survey. CSUR **56** (2024)
7. Daskin, M.: Network and discrete location: models, algorithms and applications. Journal of the Operational Research Society **48**, 763–764 (1997)
8. Davis, R.A., Lii, K.S., Politis, D.N.: Remarks on some nonparametric estimates of a density function. In: Selected Works of Murray Rosenblatt, pp. 95–100. Springer (2011)
9. Fragkathoulas, C., Papanikou, V., Karidi, D.P., Pitoura, E.: On explaining unfairness: An overview. In: 2024 IEEE 40th International Conference on Data Engineering Workshops (ICDEW). pp. 226–236. IEEE (2024)
10. Fragkathoulas, C., Pitoura, E.: Ugce: User-guided incremental counterfactual exploration. arXiv preprint:2505.21330 (2025)
11. Friedler, S.A., Scheidegger, C., Venkatasubramanian, S.: The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making. Communications of the ACM **64**, 136–143 (2021)
12. Goethals, S., Martens, D., Calders, T.: Precof: counterfactual explanations for fairness. Machine Learning **113**(5), 3111–3142 (2024)
13. Gonzalez, T.F.: Clustering to minimize the maximum intercluster distance. Theoretical computer science **38**, 293–306 (1985)
14. Guidotti, R.: Counterfactual explanations and how to find them: literature review and benchmarking. Data Mining and Knowledge Discovery **38**, 2770–2824 (2024)
15. Guidotti, R., Monreale, A., Giannotti, F., Pedreschi, D., Ruggieri, S., Turini, F.: Factual and counterfactual explanations for black box decision making. IEEE Intelligent Systems **34**, 14–23 (2019)
16. Guyomard, V., Fessant, F., Guyet, T., Bouadi, T., Termier, A.: Generating robust counterfactual explanations. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 394–409. Springer (2023)
17. Hochba, D.S.: Approximation algorithms for np-hard problems. ACM Sigact News **28**, 40–52 (1997)
18. Kanamori, K., Takagi, T., Kobayashi, K., Arimura, H.: Dace: Distribution-aware counterfactual explanation by mixed-integer linear optimization. In: IJCAI. pp. 2855–2862 (2020)
19. Kanamori, K., Takagi, T., Kobayashi, K., Ike, Y.: Counterfactual explanation trees: Transparent and consistent actionable recourse with decision trees. In: International Conference on Artificial Intelligence and Statistics. pp. 1846–1870. PMLR (2022)
20. Kavouras, L., Tsopelas, K., Giannopoulos, G., Sacharidis, D., Psaroudaki, E., Theologitis, N., Rontogiannis, D., Fotakis, D., Emiris, I.: Fairness aware counterfactuals for subgroups. Advances in Neural Information Processing Systems **36**, 58246–58276 (2023)
21. Kuratomi, A., Lee, Z., Tsaparas, P., Junior, G.D., Pitoura, E., Lindgren, T., Papapetrou, P.: Counterfair: Group counterfactuals for bias detection, mitigation and subgroup identification. In: 2024 IEEE International Conference on Data Mining (ICDM). pp. 181–190. IEEE (2024)
22. Kuratomi, A., Pitoura, E., Papapetrou, P., Lindgren, T., Tsaparas, P.: Measuring the burden of (un) fairness using counterfactuals. In: Joint European conference on machine learning and knowledge discovery in databases. pp. 402–417. Springer (2022)
23. Ley, D., Mishra, S., Magazzeni, D.: Globe-ce: A translation based approach for global counterfactual explanations. In: International conference on machine learning. pp. 19315–19342. PMLR (2023)

24. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM computing surveys (CSUR) **54**, 1–35 (2021)
25. Mothilal, R.K., Sharma, A., Tan, C.: Explaining machine learning classifiers through diverse counterfactual explanations. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. pp. 607–617 (2020)
26. Pitoura, E., Stefanidis, K., Koutrika, G.: Fairness in rankings and recommendations: an overview. The VLDB Journal pp. 1–28 (2022)
27. Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., De Bie, T., Flach, P.: Face: feasible and actionable counterfactual explanations. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. pp. 344–350 (2020)
28. Rawal, K., Lakkaraju, H.: Beyond individualized recourse: Interpretable and interactive summaries of actionable recourses. Advances in Neural Information Processing Systems **33**, 12187–12198 (2020)
29. Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144 (2016)
30. Russell, C.: Efficient search for diverse coherent explanations. In: Proceedings of the conference on fairness, accountability, and transparency. pp. 20–28 (2019)
31. Sharma, S., Henderson, J., Ghosh, J.: Certifai: A common framework to provide explanations and analyse the fairness and robustness of black-box models. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. pp. 166–172 (2020)
32. Tansel, B.C., Francis, R.L., Lowe, T.J.: State of the art—location on networks: a survey. part i: the p-center and p-median problems. Management science **29**, 482–497 (1983)
33. Ustun, B., Spangher, A., Liu, Y.: Actionable recourse in linear classification. In: Proceedings of the conference on fairness, accountability, and transparency. pp. 10–19 (2019)
34. Verma, S., Boonsanong, V., Hoang, M., Hines, K., Dickerson, J., Shah, C.: Counterfactual explanations and algorithmic recourses for machine learning: A review. ACM Computing Surveys **56**, 1–42 (2024)
35. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the gdpr. Harv. JL & Tech. **31**, 841 (2017)
36. Verma, S., Rubin, J.: Fairness definitions explained. In: FairWare@ICSE (2018)
37. Ding, F., Hardt, M., Miller, J., Schmidt, L.: Retiring adult: New datasets for fair machine learning. Advances in neural information processing systems **34**, 6478–6490 (2021)
38. Talbot, J., Lin, S., Hanrahan, P.: An extension of wilkinson's algorithm for positioning tick labels on axes. IEEE Transactions on visualization and computer graphics **16**, 1036–1043 (2010)
39. Mitchell, S., OSullivan, M., Dunning, I.: Pulp: a linear programming toolkit for python. The University of Auckland, Auckland, New Zealand **65**, 25 (2011)
40. Scott, D.W.: Multivariate density estimation: theory, practice, and visualization. John Wiley & Sons (2015)
41. Le Quy, T., Roy, A., Iosifidis, V., Zhang, W., Ntoutsi, E.: A survey on datasets for fairness-aware machine learning. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **12**, e1452 (2022)