Causally Perturbed Fairness Testing

CHENGWEN DU, IDEAS Lab, University of Birmingham, UK
TAO CHEN*, IDEAS Lab, University of Birmingham, UK

To mitigate unfair and unethical discrimination over sensitive features (e.g., gender, age, or race), fairness testing plays an integral role in engineering systems that leverage AI models to handle tabular data. A key challenge therein is how to effectively reveal fairness bugs under an intractable sample size using perturbation. Much current work has been focusing on designing the test sample generators, ignoring the valuable knowledge about data characteristics that can help guide the perturbation and hence limiting their full potential. In this paper, we seek to bridge such a gap by proposing a generic framework of causally perturbed fairness testing, dubbed `CausalFT`. Through causal inference, the key idea of `CausalFT` is to extract the most directly and causally relevant non-sensitive feature to its sensitive counterpart, which can jointly influence the prediction of the label. Such a causal relationship is then seamlessly injected into the perturbation to guide a test sample generator. Unlike existing generator-level work, `CausalFT` serves as a higher-level framework that can be paired with diverse base generators. Extensive experiments on 1296 cases confirm that `CausalFT` can considerably improve arbitrary base generators in revealing fairness bugs over 93% of the cases with acceptable extra runtime overhead. Compared with a state-of-the-art approach that ranks the non-sensitive features solely based on correlation, `CausalFT` performs significantly better on 64% cases while being much more efficient. Further, `CausalFT` can better improve bias resilience in nearly all cases.

CCS Concepts: • **Software and its engineering → Search-based software engineering**; **Software creation and management**; • **Computing methodologies → Machine learning**.

Additional Key Words and Phrases: Fairness Testing, AI/DNN Testing, Software Engineering for AI

ACM Reference Format:
Chengwen Du and Tao Chen. 2025. Causally Perturbed Fairness Testing. *ACM Trans. Softw. Eng. Methodol.* 1, 1, Article 1 (January 2025), 35 pages.

1 Introduction

In 2019, it was reported that the learned model, used by US hospitals to predict which patients needed extra medical care, always decided that black patients are more likely to pay for active interventions like emergency hospital visits—despite showing signs of uncontrolled illnesses [52]. Similarly, it has been revealed that 75% of the current sentiment analysis models show clear bias against women, exacerbating social inequity [26, 42]. Indeed, many software systems are increasingly leveraging the power of AI techniques, including machine/deep learning models, for making data-driven predictions and decisions in various domains with a large amount of tabular data, such as health care, criminal justice, civil service [11, 15, 62], and software engineering [34–36, 67]. However, as shown in the above examples, an undesired property of AI systems is that they can make discriminatory, biased, and unfair predictions, leading to severe societal impacts [23].

*Corresponding author: Tao Chen, t.chen@bham.ac.uk.

Authors' Contact Information: Chengwen Du, cxd394@student.bham.ac.uk, IDEAS Lab, University of Birmingham, Birmingham, UK; Tao Chen, t.chen@bham.ac.uk, IDEAS Lab, University of Birmingham, Birmingham, UK.

ACM 1557-7392/2025/1-ART1
https://doi.org/

Discrimination is often defined with respect to some sensitive features, such as age, race, gender, etc, against other non-sensitive counterparts [9, 32, 33]. What features are sensitive is domain-dependent and it is known in advance [79]. Intuitively, discrimination (or fairness bug) happens when an AI system/model tends to make different decisions/predictions for distinct individuals (individual fairness [83]) or subgroups (group fairness [76]), which only differ on the values of the sensitive feature(s).

The fairness bugs related to unwanted discrimination are deeply hidden in AI systems. The reasons for that can be vast, for example, the data samples used to train the AI model might be unfair and biased in the first place, or the model's hyperparameters have not been well-tuned [16, 17, 20, 22, 69] to mitigate fairness bugs [25]. All of these factors can unintentionally introduce fairness bugs in the trained AI system. However, finding those fairness bugs is complex, because (1) the full explainability of the AI model remains an open problem [79]; and (2) the possible input samples that the AI model can take are often intractable. For instance, the Kdd dataset has 19 features with 4.13×10^{15} possible samples [3].

To that end, fairness testing becomes a crucial step for engineering AI systems that process tabular data. The goal is to find as many individual discriminatory instances[1] as possible [9, 77, 79], each of which represents a fairness bug that demonstrates the existence of discrimination. Indeed, over the past few years, several test sample generators, including white-box [77, 79, 83] and black-box [9, 31, 39] ones, have been proposed for automatically perturbing the testing data, generating new, unforeseen test samples that can reveal hidden fairness bugs in AI systems. Yet, despite the recent advancements in fairness testing, existing generators have mostly relied on purely randomized perturbation around (or guided by) the sensitive feature, i.e., randomly searching and exploring the sample space without additional guidance and information other than the known sensitive feature. As such, those generators have not considered the interrelations between sensitive features and the other non-sensitive counterparts, and how they can help the perturbation in fairness testing. Indeed, prior works, such as PC fairness [66], have explored the interrelations between sensitive and non-sensitive features in fairness analysis. However, they mostly focus on fairness measurement rather than test case generation for fairness testing of AI systems. Given the large search space to be perturbed in the fairness testing and the confirmed fact that some non-sensitive features can interact with the sensitive one to jointly influence the fairness of the AI model [23, 80], missing such valuable information can inevitably limit the generator in finding fairness bugs, leaving its full potential untapped.

In this paper, we propose `CausalFT`, a generic framework for more effectively finding fairness bugs under the concept of causally perturbed fairness testing for AI systems that handle tabular data. The key novelty/idea is to extract the causal relationships between sensitive features and their non-sensitive counterparts, which are then injected into the perturbation to guide the test sample generation. What makes `CausalFT` unique is that, unlike most existing work that focuses on the test sample generator [9, 31, 39, 77, 79, 83], `CausalFT` serves as a higher-level, generator-agnostic framework that can be seamlessly paired with different base generators to enable causal perturbation, including white- and black-box ones. Notably, our contributions are:

- By leveraging the notion of causal inference, we build a causal graph from the training data and extract all directly and causally relevant non-sensitive features to the concerned sensitive feature that affects the class label.
- We then compute the causal effect between all non-sensitive features identified and the concerned sensitive feature, from which we select the non-sensitive feature with the highest

[1]In this work, we use instance and sample interchangeably.

causal effects, i.e., the one that is the most causally relevant to the sensitive feature and jointly influences the class label.

- Agnostic to the test sample generator, we inject the extracted causal relationship into a given generator for guiding how to perturb, forming a transformed causal perturbation, in two ways:
  - The definition of individual discriminatory instances is relaxed to consider both the concerned sensitive feature and its most causally relevant non-sensitive counterpart as the source of discrimination rather than the concerned sensitive feature only.
  - The perturbation takes place on the features other than the identified sensitive feature and the non-sensitive counterpart.
- `CausalFT` is experimentally evaluated on eight datasets with 2-3 sensitive features, across four AI models, six generators, and three fairness metrics, leading to 1296 cases of investigation.

The evaluation reveals encouraging results: with acceptable extra overhead ($\approx$ 270 seconds in the worst case), `CausalFT` considerably improves a given generator in 1209 out of the 1296 cases (93%), finding up to 420 more fairness bugs; it also performs much better than correlation-ranked non-sensitive features over 831 out of the 1296 cases (64%), together with 34% cases of similar results, while making the AI system/model more robust to bias in nearly all cases.

To promote open science, all artifacts can be found at our repository: https://github.com/ideas-labo/causalft.

The remainder of the paper is organized as below: Section 2 introduces the preliminaries and the observations that motivate our work. Section 3 elaborates on the designs of `CausalFT`. Section 4 presents the experiment setup and Section 5 analyzes the results, followed by a discussion in Section 6. Sections 7, 8, and 9 present the threats to validity, related work, and conclusion, respectively.

2 Preliminaries

2.1 Model Fairness

A fairness bug refers to any imperfection in an AI system that causes a discordance between the existing and required fairness conditions [23], which is relevant to two concepts:

- **Sensitive feature:** The feature in the dataset that is known to be legally or ethically protected as it could influence outcomes in a way that leads to discrimination [70]. Common examples are *Gender*, *Age*, and *Race*.
- **Non-sensitive feature:** All remaining unprotected features in the dataset would be non-sensitive features [76].

Fundamentally, an AI model could have a fairness issue if its prediction outcome could be solely influenced by different values of the sensitive feature. For example, assuming that an AI model is trained to predict whether a person has income above $50K per year using features, such as $\{Age, Relationship, Workclass, Occupation\}$. Suppose that, in the discretized representation, a sample can be represented as $\{23, 1, 3, 7\}$ for which the model predicts the label as false. Now, if the model predicts true for a new input sample, i.e., $\{33, 1, 3, 7\}$, created from the original one by only changing the *Age* from 23 to 33, then it means there is an unfair discrimination. Here, we say the sample $\{23, 1, 3, 7\}$ is an **individual discriminatory instance**, which reflects a fairness bug.

2.2 Fairness Testing and Problem Formulation

Given a trained model, the problem of fairness testing is to generate new, unforeseen individual discriminatory instances for a concerned sensitive feature[2] that can reveal fairness bugs in the AI system under test, deriving from some testing data. This is challenging because there are too many possible samples to explore, e.g., for the datasets considered in this work, the size of unique samples ranges from 9.2×10^6 to 4.13×10^{15}.

As such, an automatic test generator is often a search algorithm as part of the field of Search-based Software Engineering [21, 44], wherein the key is to pair two samples, from some testing data, that only differ on the concerned sensitive feature and randomly perturb all non-sensitive features (or sometimes any features) on both samples as guided by some fitness, hoping that such a perturbation can find more real individual discriminatory instances that would actually cause the trained model to predict the different outcome when changing solely on the sensitive feature. For example, suppose that the feature at index 0 is the concerned sensitive feature. Then, the initial sample pair $\mathbf{x}_a$ and $\mathbf{x}_b$ on the left might lead to the pair on the right after the perturbation:

$$\begin{aligned} \mathbf{x}_a &= \{0, 7, 4, 5, 1\} \Longrightarrow \mathbf{x'}_a = \{0, 2, 8, 5, 3\} \\ \mathbf{x}_b &= \{1, 7, 4, 5, 1\} \Longrightarrow \mathbf{x'}_b = \{1, 2, 8, 5, 3\} \end{aligned} \tag{1}$$

If both pairs can cause the trained model to generate different outcomes, then we have four individual discriminatory instances. Our goal of fairness testing here is to find as many (unique) individual discriminatory instances as possible.

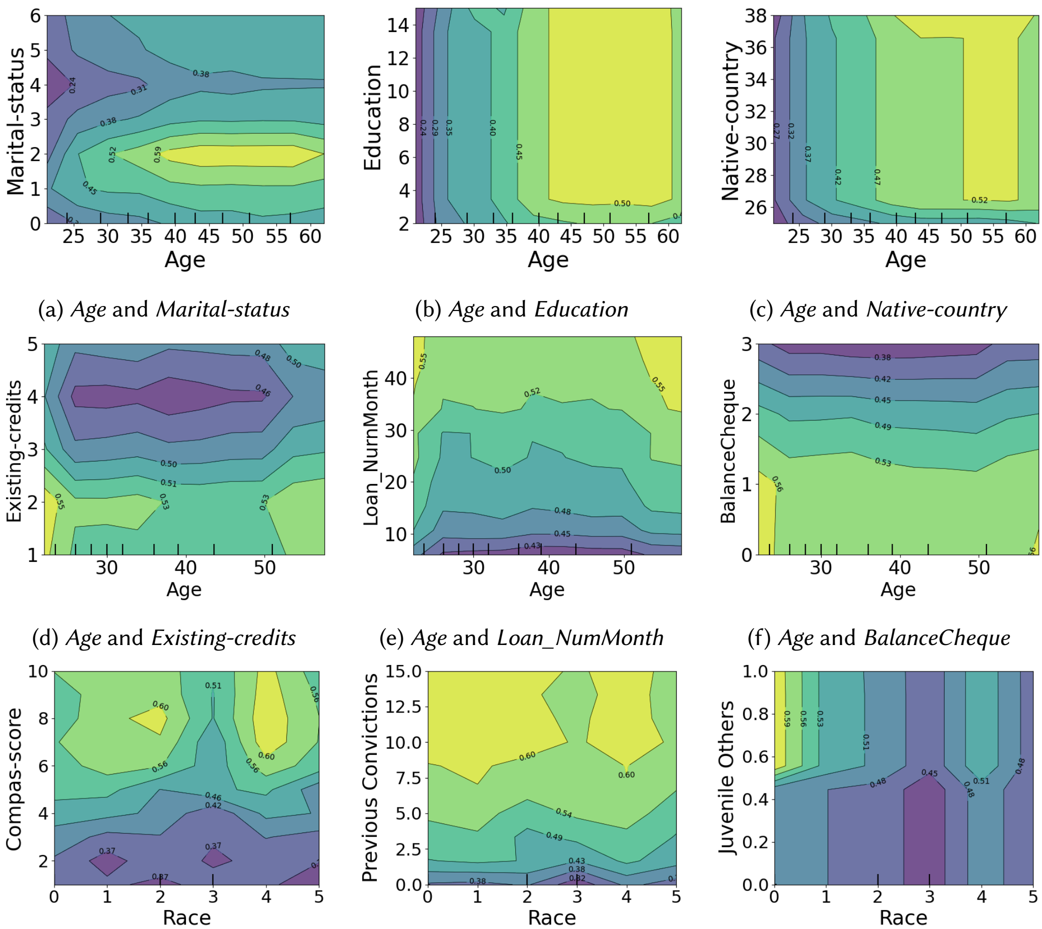
Two types of metrics exist to evaluate the capability of test generators in finding the fairness bugs in a trained model: individual fairness [29, 83] and group fairness [27, 40, 46, 48, 51, 58, 59, 76]. The former is mainly relevant to the number of individual discriminatory instances in the generated samples; while the latter means a fair outcome demands the existence of parity between different feature groups by, e.g., *gender* or *race*. It has been shown that individual fairness can be correlated with the group fairness counterpart [13, 43].

2.3 Observations and Motivation

To understand how discrimination instances can easily cause AI systems to produce unfair outcomes, we examine the trained AI model and make predictions over several real-world tabular datasets. We randomly generate samples to test an AI system that contains a Deep Neural Network (DNN), and identify those that can reveal fairness bugs. As shown in Figure 1, we observe the following patterns:

- **Observation 1:** For a given sensitive feature, there exists a non-sensitive feature that is far more commonly interacting with it to create discrimination instances. For example, when the *Age* is the sensitive feature for the Adult dataset, we see that there are more complex interactions between *Age* and *Marital-status*, which affect the decisions made by the model. In contrast, other non-sensitive features like *Education* are more independent of the changes of *Age* when influencing the AI model's prediction, as the probability boundary largely varies according to the values of *Education* but not *Age*.
- **Observation 2:** The non-sensitive feature that causes significant interactions with the sensitive counterpart to influence the prediction is uncertain. For example, even for the same sensitive feature *Age*, it can create considerable interactions with varying non-sensitive features on Adult and German datasets, even though they share many common features.

[2]In fairness testing, there is often only one concerned sensitive feature each time even if multiple sensitive features exist in the dataset [9, 77, 79].



(a) *Age* and *Marital-status* (b) *Age* and *Education* (c) *Age* and *Native-country*

(d) *Age* and *Existing-credits* (e) *Age* and *Loan_NumMonth* (f) *Age* and *BalanceCheque*

(g) *Race* and *Compas-score* (h) *Race* and *Pre. Convictions* (i) *Race* and *Juvenile Others*

Fig. 1. Exampled partial dependence plots for the most significant interactions between the sensitive and non-sensitive features in DNN prediction. (a), (b), and (c) are the ADULT dataset when *Age* is the sensitive feature; (d), (e), and (f) are the GERMAN dataset when *Age* is the sensitive feature; (g), (h), and (i) are the COMPAS dataset when *Race* is the sensitive feature. The numbers in the figures denote the inferred probability belongs to the positive class.

The above makes sense in the real world, because, for example, in the ADULT dataset, *Age* can more considerably interact with the *Marital-status* to determine the AI model's prediction: younger individuals are more likely to be single or in non-married relationships while older individuals are more likely to have other marital statuses [12].

Those observations lead to the following key insight when testing the fairness of AI systems:

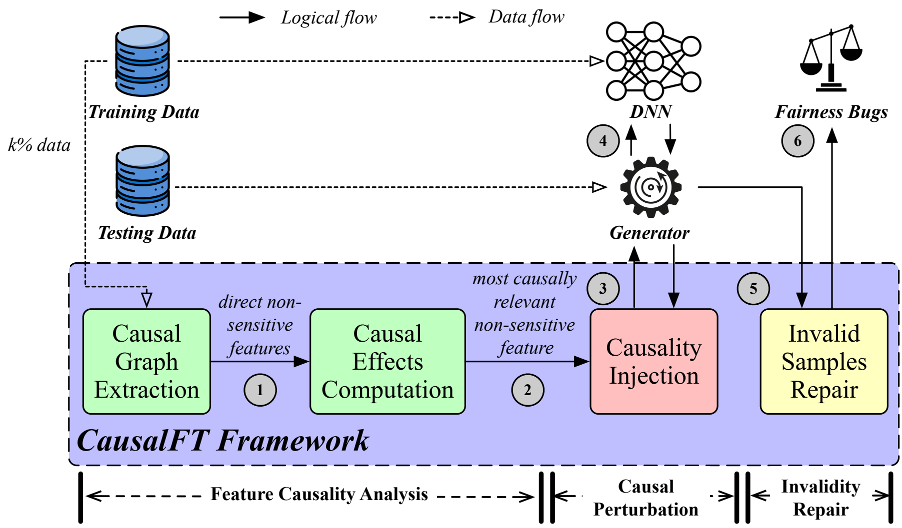


Fig. 2. Workflow of CausalFT for fairness testing.

Key Insight

The strong interactions between a sensitive feature and its non-sensitive counterpart should be explicitly handled when testing the fairness of AI systems, as they serve as the key to determining model prediction, leading to discriminatory instances.

However, existing generators for fairness testing have failed to handle the above characteristics, since they typically perturb features in a random or heuristic manner without considering the interactions between a sensitive feature and its non-sensitive counterparts [9, 31, 33, 39, 63, 77, 79, 83]. As such, existing generators generally treat all non-sensitive features as equally important in the perturbation. For example, when *Age* is the sensitive feature for the Adult dataset, perturbing both *Marital-status* and *Education* with equal chance is clearly not ideal, as the former can more significantly interact with *Age* to change the prediction outcome. This can cause several devastating consequences:

- ineffective exploration of the most probable bias-induced regions in the input space;
- weak exploitation of the found fairness bugs, as the samples in the generated pairs are never re-paired with the other samples;
- wasting computational resources since those less influential non-sensitive features with limited interaction with the sensitive one can consume the majority of the computation.

All of the above motivate our work: we seek to overcome those limitations of AI system fairness testing by explicitly taking the newly discovered observations into account.

3 Fairness Testing with Causal Perturbation

Since the perturbation is crucial for test sample generation in fairness testing, the ultimate goal of `CausalFT` is to improve the effectiveness of perturbation by exploiting the learned causal relationships between sensitive and non-sensitive features. The key idea is that diversifying those non-sensitive features, which are more causally correlated with the concerned sensitive ones, as the starting point of the perturbation while keeping such diversity would be more likely to help find individual discriminatory instances that reveal fairness bugs. This is because those non-sensitive features and the sensitive counterpart have a larger chance to jointly influence the prediction of an AI system/model, and hence it is important to keep them diverse without being affected too much by the perturbation. `CausalFT` aims to identify those non-sensitive features and explicitly handle them during perturbation.

As with existing work [9, 77, 79], we assume that there is only one concerned sensitive feature each time, known in advance. The benchmark dataset is split into training and testing data. The former trains an AI model and is used by the causal analysis in `CausalFT` while the latter is used in the perturbation/generation process. We follow the "rule-of-thumb" that 70% of a dataset is used for training while the remaining 30% is used to seed the generator [53]. From Figure 2 and Algorithm 1, `CausalFT` has three key phases:

- **Feature Causality Analysis** (lines 1-3): Here, we seek to analyze the causal relationships between the concerned sensitive and other non-sensitive features, from which we pick the most causally relevant one to serve as part of the "synthetic sensitive set" along with the sensitive one.
- **Causal Perturbation** (lines 4-13): We inject the most causally relevant non-sensitive feature selected into the existing perturbation of a base generator. This transforms the generator in a way that its perturbation is guided by causal knowledge, diversifying the identified non-sensitive feature and its sensitive counterpart while reducing the space of perturbation.
- **Invalidity Repair** (lines 14-16): When possible, we find new seeded samples from the testing data to pair with samples in an invalid pair.

As such, `CausalFT` serves as a generic, generator-agnostic framework that can be seamlessly paired with different perturbation-based generators for fairness testing without requiring substantial modification. In what follows, we will articulate the key designs behind `CausalFT` in detail.

3.1 Feature Causality Analysis

3.1.1 Causal Graph Extraction. Using k% of the training data (default to $k = 100$), a major step in `CausalFT` is to build a causal model therein for fairness testing, which can be represented as a directed acyclic graph (DAG) where each node can be a feature (f_i) or label (y) while a path $f_i \rightarrow f_j$ is the causal implication from f_i to f_j. Since the causal relationship is unidirectional, we fix a concerned sensitive feature (f_s) as the starting node and the label as the ending node, but we permit several possible intermediate features within their causal path, e.g., $f_s \rightarrow f_i \rightarrow y$. All those paths will form a DAG, and our goal is to identify which are the non-sensitive features that are causally affected by changing the value of a sensitive feature, which would then influence the label in the dataset. It is worth noting that, for a dataset, we observe minimal change in the obtained DAG even when using different sensitive features as the starting node. Indeed, there are several methods/models that can be used for the above purpose. In `CausalFT`, we leverage LiNGAM [57] to build the causal graph[3], because:

- It has proven effectiveness for a wide range of tasks [56].
- It is computationally efficient due to its linear and non-Gaussian model. Indeed, this might not generalize to all cases, but we found that it fits the datasets studied well.
- It also contains an effective pruning mechanism that reduces the complexity of the causal graph produced.

[3]We used LiNGAM under label encoding of our discrete data, for which it has been proved to preserve the same assumption [57].

Algorithm 1: `CausalFT` framework

Input: Training/testing data $\mathcal{D}_{train}/\mathcal{D}_{test}$; concerned sensitive feature f_s; generator $\mathcal{G}$; generated size s; $k\%$ of $\mathcal{D}_{train}$; the AI model under test $\mathcal{T}$.

Declare: Causal graph $\mathcal{M}$; selected non-sensitive feature f_c.

Output: A set of generated individual discriminatory instances $\mathcal{S}$.

/* Feature causality analysis. */

$\mathcal{M}$, $\{f_1, ..., f_l\} \leftarrow$ doCausalityAnalysis($k\% \times \mathcal{D}_{train}$,$f_s$)

$\{\{f_1, c_1\}, ..., \{f_1, , c_l\}\} \leftarrow$ getCausalEffect($\{f_1, ..., f_l\}$,$\mathcal{M}$)

f_c = the direct non-sensitive feature with the largest causal effect

/* Causal perturbation; this might slightly differ according to the base generator. */

while $|\mathcal{S}| < s$ **do**

  $\{\mathbf{x}_a,\mathbf{x}_b\} \leftarrow$ randomly get a sample pair from $\mathcal{D}_{test}$

  if $\mathbf{x}_a$ *and* $\mathbf{x}_b$ *only differ on* $\{f_s, f_c\}$ *while their predictions under* $\mathcal{T}$ *are different* **then**

    $\mathcal{S} = \mathcal{S} \cup$ the new unique sample(s) from $\{\mathbf{x}_a,\mathbf{x}_b\}$

  else

    $\mathbf{x}_b \leftarrow$ a random sample differed from $\mathbf{x}_a$ only on $\{f_s, f_c\}$

    $\{\mathbf{x}'_a,\mathbf{x}'_b\} \leftarrow$ perturbWithDNN($\mathcal{T}$,$\mathcal{G}$,f_s,f_c,$\mathbf{x}_a$,$\mathbf{x}_b$)

    $\mathcal{S} = \mathcal{S} \cup$ the new unique sample(s) from $\{\mathbf{x}'_a,\mathbf{x}'_b\}$

    /* Note that only a perturbed sample and a sample from $\mathcal{D}_{test}$ can be paired. */

    if *neither* $\mathbf{x}'_a$ *nor* $\mathbf{x}'_b$ *belongs to* $\mathcal{D}_{test}$ **then**

      $\{\mathbf{x}'_a,\mathbf{x}''_a\}$, $\{\mathbf{x}'_b,\mathbf{x}''_b\} \leftarrow$ find other sample(s) in $\mathcal{D}_{test}$, if any, that can be paired with $\mathbf{x}'_a$ and/or $\mathbf{x}'_b$ according to the true IDI definition

    end

  end

end

/* Invalidity repair. */

for $\forall\{\mathbf{x}'_a,\mathbf{x}'_b\} \in \mathcal{S}$ *that is invalid under the true definition* **do**

  $\{\mathbf{x}'_a,\mathbf{x}''_a\}$, $\{\mathbf{x}'_b,\mathbf{x}''_b\} \leftarrow$ find other sample(s) in $\mathcal{D}_{test}$, if any, that make $\mathbf{x}'_a$ and/or $\mathbf{x}'_b$ valid under the true definition

end

return $\mathcal{S}$

In a nutshell, LiNGAM only needs to take the dataset, starting node, and ending node as inputs, and it then computes the independent components involved by using an ICA model [30]. The output is a linear, non-Gaussian, acyclic model that contains the causal path from the starting node to the ending one, including any intermediate nodes.

Drawing on the produced/verified causal graph from LiNGAM, `CausalFT` takes two more steps:

(1) Extract all causal paths starting from the concerned sensitive feature and ending at the label, e.g., $f_s \rightarrow f_i \rightarrow y$.
(2) Within those causal paths, identify all (unique) non-sensitive features that have a **direct** causal relationship to the sensitive one. For example, if we have a causal path $f_s \rightarrow f_i \rightarrow f_j \rightarrow y$, then only f_i is of interest as f_j exhibits an indirect relationship to f_s.

We employ LiNGAM to construct a directed acyclic causal graph (DAG) and identify direct edges from the sensitive attribute to non-sensitive features. Only features with direct causal connections are selected for perturbation, ensuring causal validity. As shown in Figure 3, several causal paths can be extracted, such as *Age* $\rightarrow$ *Marital-status* $\rightarrow$ *Gender* $\rightarrow$ *Class-label* and *Age* $\rightarrow$ *Marital-status* $\rightarrow$ *Class-label*. Thus, *Age* has a direct causal edge to *Marital-status*, while its influence on *Gender* is indirect via *Marital-status*. It is worth noting that the method that builds the causal graph in `CausalFT` can be any off-the-shelf algorithm, hence LiNGAM can be easily replaced with an alternative solution when needed.

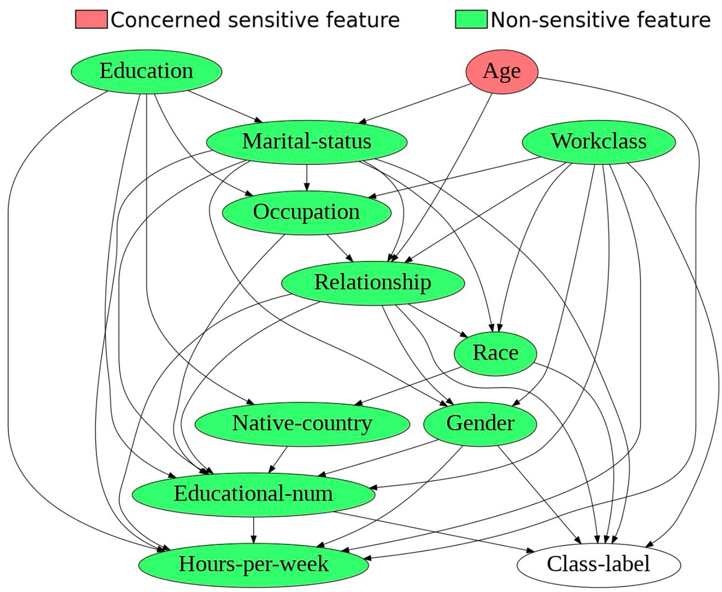


Fig. 3. The causal graph built for the Adult dataset (*Age* being the only concerned sensitive feature).

To validate the robustness of the causal graphs generated by LiNGAM, we compare it against other causal discovery algorithms, PC [66] and GES [24], on a bootstrap stability analysis using the `Tetrad` package with 1000 bootstraps. We compute the hamming distance on the edge's adjacency matrix of every pair of the graph generated and report the average distance as a measure of stability. We found that:

- LiNGAM has 1.6 edge differences.
- PC has 2.1 edge differences.
- GES has 2.6 edge differences.

These results confirm that LiNGAM generates robust and reliable causal structures for `CausalFT`.

3.1.2 Causal Effects Computation. To compute the causal effects between the concerned sensitive feature f_s and each of its direct causally relevant non-sensitive features f_n as extracted from the causal graph, `CausalFT` follows the steps below:

(1) Collect m random samples from the $k\%$ training data ($m = 100$ in this work[4]).
(2) Feed the m samples as inputs into the causal graph built to estimate the probabilities of f_s and f_n on affecting the label y under their possible values sampled.
(3) Calculate the causal effect from f_s to f_n on the class label using the *do-calculus* of counterfactual cases via:

$$c(f_s, f_n) = \frac{\theta}{n^2} \sum_{\alpha \in V_\alpha} \sum_{\beta \in V_\beta} |p(y|do(f_s = \alpha)) - p(y|do(f_n = \beta))| \quad (2)$$

[4]Note that m differs from the k used to train/build the causal graph, but for computing the causal effects based on a given graph. Increasing or setting m as $k\%$ of the training data would lead to little gains due to the bootstrapping, while the computational overhead increases exceptionally.

where α and β are two values for f_s and f_n respectively, each chosen from their permitted set V_α and V_β; θ is the fitted coefficient between f_s and f_n from the causal graph, which represents the strength of the causal relationship between them.
(4) Repeat from 1) for 20 runs following a bootstrapping with replacement.
(5) Compute the median of the causal effect $c(f_s, f_n)$ over the 20 repeats, denoted as c_n.

From the above, we can then rank every non-sensitive feature that is directly and causally relevant to the sensitive counterpart using c_n; the feature f_c with the highest c_n is what we are seeking. Indeed, it is possible to consider more than one non-sensitive feature. However, we use the most causally relevant non-sensitive feature from those directly related ones in `CausalFT` because:

- We discovered that there is always one non-sensitive feature with a significantly higher c_n than the others.
- Including more non-sensitive features would have the risk of not being able to explore the search space sufficiently while producing too many invalid samples, as we will discuss in Section 5.3.
- If a non-sensitive feature is not causally relevant to the sensitive counterpart, then it makes less sense to correlate them, which might lead to misleading perturbation.

As a result, we consider only the most causally relevant non-sensitive feature to be a known sensitive one. For example, under the Law School dataset, when *Race* is the concerned sensitive feature, among all its directly and causally relevant non-sensitive features, *Lsat* has the highest causal effect of 4.75, which is significantly higher than the 2.29 of the second highest feature *Decile3* following by the 2.11 of *Decile1b*.

3.2 Causal Perturbation

3.2.1 Common Parts of Perturbation. Since current test sample generators essentially resemble a search process for finding individual discriminatory instances via perturbing the samples in the testing dataset, there are two important designs therein:

(1) **How to define individual discriminatory instances in the perturbation?** This directly influences how a pair can be formed and the direction of the perturbation.
(2) **What features to perturb?** For each pair of samples, this determines the features to be changed during perturbation[5], which underpins the search space and hence has a major impact on the success of the perturbation.

While distinct generators can be influenced by the above designs to different extents, they generally do not consider the usefulness of the non-sensitive feature and its causal relationship to the concerned sensitive counterpart during perturbation at all. By using the most causally relevant non-sensitive feature identified from the causal analysis in Section 3.1, we seek to transform the perturbation in a given generator to make it causality aware—a key contribution of this work.

3.2.2 Seamless Causality Injection. In `CausalFT`, we inject causality into the above two design aspects to transform the perturbation. Specifically, existing generators use the **true definition** of individual discriminatory instance to influence the perturbation, which only considers the sensitive feature(s) when determining **how to define individual discriminatory instances in the perturbation**, i.e., $\mathbf{x}_a$ is an individual discriminatory instance in a pair if there is another sample $\mathbf{x}_b$ such that they differ on the concerned sensitive feature f_s but all other feature values are the same while their predictions by the tested AI model (y_a and y_b) are different. That is:

$$f_{s,a} \neq f_{s,b} \text{ and } \forall f_{i,a} = f_{i,b} : f_{i,a}, f_{i,b} \in \mathcal{F}' \text{ while } y_a \neq y_b \tag{3}$$

[5]Note that perturbation changes one feature each time and often it changes the same feature of both samples in the pair to a different value.

whereby $f_{s,a}$ and $f_{s,b}$ are the values of the concerned sensitive feature for the two samples, respectively; similarly, $f_{i,a}$ and $f_{i,a}$ are the values of any remaining non-sensitive features in those samples from the set $\mathcal{F}'$. As such, for each sample $\mathbf{x}_a$ that is not yet an individual discriminatory instance from the testing data, according to the true definition of individual discriminatory instance, those generators initially build another sample $\mathbf{x}_b$ to form a pair for the perturbation.

To inject causality into the definition of individual discriminatory instance that influences the perturbation, instead, we temporarily append the direct and most causally relevant non-sensitive feature f_c along with the concerned sensitive one f_s as a new form of "synthetic sensitive set", creating a **relaxed definition** of the individual discriminatory instance. That is, not one but a pair of features would be considered in the sensitive part, which then creates a more diverse sample $\mathbf{x}_b$ that might be different from $\mathbf{x}_a$ on both f_c and f_s for the perturbation and determining when the perturbation of a pair should terminate. Formally, this means that:

$$f_{s,a} \neq f_{s,b} \text{ or } f_{c,a} \neq f_{c,b} \text{ and } \forall f_{i,a} = f_{i,b} : f_{i,a}, f_{i,b} \in \mathcal{F}'' \text{ while } y_a \neq y_b \tag{4}$$

where $\mathcal{F}''$ is the feature set after ruling out f_s and f_c; $f_{c,a}$ and $f_{c,b}$ are the values of the most causally relevant non-sensitive feature for the sensitive counterpart under the two samples, respectively. Clearly, a pair that meets the true definition of individual discriminatory instance would certainly meet the relaxed definition, but the reverse might not be true. Thus, pairs that are invalid under the true definition of individual discriminatory instance could still meet the relaxed definition. However, our goal is merely to diversify the perturbation of the most causally relevant sensitive and non-sensitive features, and it is easy to repair this in the end as we will show.

Example: Consider a case that *Age* is the concerned sensitive feature at the index of 5, then for a sample $\mathbf{x}_a$, existing generators could randomly generate $\mathbf{x}_b$ to form a pair used for perturbation, as shown below:

$$\begin{aligned} \mathbf{x}_a &= \{0, 3, 4, 6, 3, 25, 8, 7, 1, 1, 2\} \\ \mathbf{x}_b &= \{0, 3, 4, 6, 3, 11, 8, 7, 1, 1, 2\} \end{aligned} \tag{5}$$

Clearly, they only differ in their sensitive features *Age*. Suppose that `CausalFT` has identified *Hours-per-week* (the index of 7) as the directly and causally relevant non-sensitive feature with the highest causal effects, then, in contrast, `CausalFT` might generate a pair as below:

$$\begin{aligned} \mathbf{x}_a &= \{0, 3, 4, 6, 3, 25, 8, 7, 1, 1, 2\} \\ \mathbf{x}_b &= \{0, 3, 4, 6, 3, 11, 8, 1, 1, 1, 2\} \end{aligned} \tag{6}$$

Here, both *Age* and *Hours-per-week* could be different between $\mathbf{x}_a$ and $\mathbf{x}_b$ when the perturbation starts.

For determining **what features to perturb**, current generators mostly rule out the sensitive one and randomly perturb all the remaining features (e.g.,Zhang et al. [79]), or they simply perturb all features in a random manner (e.g.,Aggarwal et al. [9]). In contrast, according to the above "synthetic sensitive set", we further reduce the size of the perturbation by fixing both the concerned sensitive feature and its most causally relevant non-sensitive one during perturbation, hence making it less randomized. This would also help to maintain the diversity with respect to those two features across the search space. Naturally, the impact of perturbation might differ based on the generator, but the basic idea remains the same. For instance, for ADF [79], causality would be injected into both the global and local perturbation while for single perturbation generators like SG [9], this only affects the perturbation once.

Example: In the existing generators, a perturbed pair $\mathbf{x}'_a$ and $\mathbf{x}'_b$ from the original $\mathbf{x}_a$ and $\mathbf{x}_b$ at Equation (5) could be:

$$\begin{aligned}\mathbf{x}'_a &= \{1, 2, 3, 7, 5, 25, 9, 9, 2, 3, 4\}\\ \mathbf{x}'_b &= \{1, 2, 3, 7, 5, 11, 9, 9, 2, 3, 4\}\end{aligned} \tag{7}$$

where all features except the sensitive feature *Age* might have been changed. In `CausalFT`, instead, the perturbed pair from Equation (6) might become:

$$\begin{aligned}\mathbf{x}'_a &= \{1, 2, 3, 7, 5, 25, 9, 7, 2, 3, 4\}\\ \mathbf{x}'_b &= \{1, 2, 3, 7, 5, 11, 9, 1, 2, 3, 4\}\end{aligned} \tag{8}$$

and it is clear that all features could have been changed except *Age* and its directly and most causally relevant non-sensitive feature with the highest causal effects *Hours-per-week*.

For each selected non-sensitive feature, perturbation is performed by randomly sampling a new value within its observed range or valid categories. No fixed increase or decrease direction is enforced, allowing diverse yet causally informed perturbations. We can do so because of the additional mechanism in `CausalFT` to ensure constraint compliance: all perturbations respect atomic constraints by restricting values to each feature's observed range or valid categories in the training dataset (e.g., *Age* can only be chosen from 17 to 90 [8]), ensuring realistic and valid perturbed samples. As for violating constraints among features that form a realistic sample, `CausalFT` naturally mitigates this by pairing each perturbed sample with another sample from the original testing data (if neither of the samples in the pair is from the testing data), which ensures compliance with the cross-feature constraints[6].

3.3 Invalidity Repair

Finally, for pairs that are valid under the relaxed definition but not the true definition, `CausalFT` performs the following:

(1) It randomly finds, from other pairs of the original seeded samples of the testing data, the sample that differs from either sample in the invalid pair by only the concerned sensitive feature. `CausalFT` additionally pairs such a sample with the corresponding sample from the invalid pair to form a new pair.
(2) If a sample in the invalid pair cannot be repaired, we simply do nothing and consider it as a failed sample[7].

In this way, we can benefit from the causally perturbed fairness testing while still ensuring that the true definition of individual discriminatory instances can be eventually met.

Example: Suppose that in the end, we have the pair $\mathbf{x}'_a$ and $\mathbf{x}'_b$ from Equation (8) which is clearly invalid under the true definition of individual discriminatory instance. In `CausalFT`, we then repair those samples by finding the samples from other pairs of the original seeded samples that differ from $\mathbf{x}'_a$ or $\mathbf{x}'_b$ only on the concerned sensitive feature *Age* at the index of 5 with different predictions by the tested AI model (y'_a and y''_a; y'_b and y''_b), forming new pairs $\mathbf{x}'_a$ and $\mathbf{x}''_a$ together with $\mathbf{x}'_b$ and $\mathbf{x}''_b$:

$$\begin{aligned}\mathbf{x}'_a &= \{1, 2, 3, 7, 5, 25, 9, 7, 2, 3, 4\};\ y'_a = \alpha\\ \mathbf{x}''_a &= \{1, 2, 3, 7, 5, 17, 9, 7, 2, 3, 4\};\ y''_a = \beta\end{aligned} \tag{9}$$

[6]Note that we do not assume any constraint with respect to the sensitive feature as those constraints themselves, even if exist, can impose unfair discrimination, e.g., in some countries, very talented students can have already graduated from the university at a very young age.
[7]We found that this is unlikely given the large number of samples generated.

$$\mathbf{x}'_b = \{1, 2, 3, 7, 5, 11, 9, 1, 2, 3, 4\};\ y'_b = \alpha$$
$$\mathbf{x}''_b = \{1, 2, 3, 7, 5, 13, 9, 1, 2, 3, 4\};\ y''_b = \beta \tag{10}$$

Through the above, CausalFT seamlessly transforms the perturbation with causality knowledge in a base generator.

Note that even if $\mathbf{x}'_a$ and $\mathbf{x}'_b$ form a valid pair initially, CausalFT still needs to verify if either $\mathbf{x}'_a$ or $\mathbf{x}'_b$ exist in the original testing data; if not, we then also try to pair each of them with one from the testing data following the true definition of individual discriminatory instance, which is similar to the case of invalidity repair. This is important to ensure the practicality of the test cases.

It is worth noting that since CausalFT always pairs one of the invalid samples from a perturbed pair with the other original seeded sample of the testing data, rather than the other perturbed pairs, the realism of the results can be improved. An invalid sample would be discarded if it cannot be paired with any original testing data samples.

4 Experiments

We use several research questions (RQs) to evaluate CausalFT:

- **RQ$_1$:** To what extent can CausalFT improve state-of-the-art fairness testing generators?
- **RQ$_2$:** How does CausalFT perform compared to the existing method that ranks the relevance of non-sensitive features via correlation analysis?
- **RQ$_3$:** What is the impact of individual discriminatory instance definition relaxation and invalidity repair?
- **RQ$_4$:** What is the fairness improvement by retraining AI model using the test samples generated by CausalFT?
- **RQ$_5$:** How efficient is CausalFT in generating the required number of unique samples?

Since CausalFT serves as a general framework, in **RQ$_1$**, we evaluate to what extent it can be beneficial to improve existing generators for fairness testing across three different AI models. **RQ$_2$** further verifies that the particular causal inference used in CausalFT can be important for any testing improvement against existing methods that capture correlation between the features. **RQ$_3$** evaluates the impact of relaxing the individual discriminatory instance (IDI) definition and performing invalidity repair on the effectiveness and validity of the generated test cases. **RQ$_4$** naturally asks how the fairness of AI system/model can be improved if it is re-trained by using the samples generated by CausalFT. Finally, **RQ$_5$** examines the efficiency of CausalFT, as it would be less sensible if any improvements require an extensive amount of overhead to achieve.

4.1 Datasets

In this work, we consider eight datasets, containing commonly used binary classification problems for fairness testing, as shown in Table 1. These datasets are chosen because:

- They come from different domains and with a diverse number of samples and search space for perturbation.
- They contain rich characteristics and diverse demographic groups (e.g., *Age* and *Race*).
- They are real-world datasets, which strengthens the practicality of the evaluation.
- They are publicly available and widely used in prior fairness testing studies [23, 54].

All datasets come with pre-defined sensitive features. For example, the Adult dataset has *Age*, *Race*, and *Gender*.

4.2 Models

We evaluate CausalFT on four widely used models:

Table 1. The real-world datatset used in fairness testing. $|f_s|$ and $|f|$ denote the possible number of sensitive features and the number of all features, respectively.

Dataset	Domain	$\lvert f_s \rvert$	$\lvert f \rvert$	Available Sample Size	Full Sample Size
Adult [8]	Finance	3	11	45, 222	4.81×10^9
Compas [6]	Criminology	2	13	6, 172	1.45×10^8
Law School [2]	Education	2	12	20, 708	9.20×10^6
Kdd [3]	Criminology	2	19	284, 556	4.13×10^{15}
Dutch [5]	Finance	2	12	60, 420	3.58×10^7
Credit [7]	Finance	3	24	30, 000	2.01×10^{12}
Crime [4]	Criminology	2	22	2, 215	4.19×10^8
German [1]	Finance	2	20	1, 000	8.85×10^9

- **Logistic Regression (LR):** As a linear and interpretable baseline, LR serves to evaluate `CausalFT`'s performance on a simple, convex model commonly used in fairness studies. We use the implementation from `scikit-learn` with L2 regularization ($C = 1.0$).
- **Random Forest (RF):** RF provides a non-linear, ensemble-based classical model, useful for evaluating `CausalFT`'s compatibility with non-differentiable architectures. We also use the `scikit-learn`'s implementation with 100 estimators, no maximum depth restriction, and default settings otherwise.
- **Six-Layer Fully Connected Deep Neural Network (DNN_6):** A six-layer fully connected neural network (a standard setting [77]) consisting of:
  - **Input layer**: Number of neurons equal to input feature dimensions.
  - **Hidden layers**: We follow five hidden layers, each with ReLU activation functions. The number of neurons per hidden layer is empirically tuned per dataset (ranging from 64 to 256) to optimize classification performance while preventing overfitting.
  - **Output layer**: A single sigmoid-activated neuron for binary classification tasks.

  We perform optimization using the `Adam` optimizer with a learning rate of 0.001 and a batch size of 128. Early stopping is employed based on validation loss with a patience of 10 epochs to avoid overfitting. We use the implementation in `TensorFlow`.
- **Five-Layer Fully Connected Deep Neural Network (DNN_5):** To examine a diverse set of model capacities and investigate robustness under different complexities, we also evaluate a 5-layer fully connected network with the following configuration[8]:
  - **Input layer**: Number of neurons equal to input feature dimensions.
  - **Hidden layers**: Five fully connected layers with sizes [256, 256, 128, 64, 32]. Each layer uses ReLU activation. In addition to the six-layer DNN, dropout layers are inserted after the first three hidden layers (with probabilities of 0.3, 0.3, and 0.2, respectively) to reduce overfitting.
  - **Output layer**: A single sigmoid-activated neuron for binary classification tasks.

  We perform optimization using the `Adam` optimizer with a learning rate of 0.001 and a batch size of 128. This model is trained for 100 epochs with L_2 regularization, using Binary Cross-Entropy Loss as the objective function. All weights are initialized using Xavier initialization to improve training stability. As before, we use the implementation in `TensorFlow`.

4.3 Base Generators

To verify the robustness and comparability of `CausalFT`, we consider a wide range of base generators, including both the white-box and black-box generators. In essence, white-box generators exploit and extract properties from the internal working mechanisms of AI model, such as the gradient and activation of the neuron, to guide the perturbation process in test sample generation.

[8] A five-layered DNN is the smallest network used in prior work [77] for fairness testing.

Most of those generators differ in terms of the properties they leverage and how the corresponding information is extracted. The black-box generators, in contrast, do not rely on information from the internal structure of AI model but are purely based on sophisticated perturbation designs. In this work, we consider three white-box generators (`ADF` [79], `EIDIG` [77], and `NeuronFair`[83]) and other three black-box ones, i.e., `SG` [9], `ExpGA` [31], and `BREAM` [39]—all are state-of-the-art approaches with diverse characteristics. To adopt the white-box generators designed for DNN on other AI models, we slightly modify them, e.g., instead of focusing on the neuron activation, they can focus on the coefficient of terms for LR and the Gini importance on RF. Notably, we ruled out certain generators [68] because their code is not executable, as such our evaluation seeks to cover representative white-/black-box base generators. We do not claim those to be exhaustive.

It is worth noting that the above generators, either white-box or black-box, leverage the true definition of the individual discriminatory instance and perform the perturbation without the sensitive features (or simply all features). For all cases, we fix a budget of $10,000$ (unique) samples to be generated, which is a standard setting in prior work [77, 79, 83].

4.4 Metrics

We evaluate `CausalFT` on both individual and group fairness. For individual fairness, we calculate the ratio between the number of unique individual discriminatory instances under the true definition (I) and the size of all of the generated unique samples (S), namely IDI ratio ($\frac{I}{S}$) [63, 79]. In fairness testing, a higher IDI ratio means more fairness bugs are found. Note that in `CausalFT`, only a perturbed sample and a sample from $\mathcal{D}_{test}$ can be paired.

For group fairness, we divide all generated test samples into different groups according to the sensitive features. For example, in the Adult dataset, when *Gender* is the concerned sensitive feature, we divide the test samples depending on whether they have 0 (*Male*) or 1 (*Female*) on *Gender*[9]. Drawing on this, we use two common metrics in the evaluation:

- **Equal Opportunity Difference (EOD) [37]** measures the extent to which the same proportion of each group divided from the sensitive feature receives a favorable outcome. Formally, it is computed as:

$$EOD = |\mathbb{E}(\hat{y} \mid f_s = \alpha, y = 1) - \mathbb{E}(\hat{y} \mid f_s = \beta, y = 1)| \tag{11}$$

  whereby $\hat{y}$ and y are the model-predicted and actual label ($y = 1$ means positive label), respectively. α and β are values that decide the group for the sensitive feature f_s. A higher EOD means finding more fairness bugs.
- **Statistical Parity Difference (SPD) [28]** measures the difference that the expected prediction from an AI model made with respect to those groups that differ on the sensitive feature. Formally, it is expressed as:

$$SPD = |\mathbb{E}(\hat{y} \mid f_s = \alpha) - \mathbb{E}(\hat{y} \mid f_s = \beta)| \tag{12}$$

  Again, a higher SPD means a test generator is capable of revealing more fairness bugs.

4.5 Testing Procedure in Experiments

The testing procedure follows standard practice for fairness testing [23, 79, 83], including the following steps:

(1) Pick a dataset and perform preprocessing.
(2) Define a concerned sensitive feature and mark all others non-sensitive.

[9]For non-binary features such as *Age*, we follow the common way to discretize the values into groups, e.g., splitting age as $[25, 60]$ and others.

Table 2. Comparing different generators with and without CausalFT over all cases using the mean (standard deviation) of individual/group fairness metrics under 10 runs. f_s denotes the concerned sensitive feature when testing DNN_5. green cells mean a generator with CausalFT leads to higher IDI ratio/EOD/SPD (hence better at finding fairness bugs) with $p < 0.05$ and non-trivial $\hat{A}_{12}$; red cells denote a generator without CausalFT can better reveal fairness bugs with $p < 0.05$ and non-trivial $\hat{A}_{12}$. Statistically insignificant comparisons (i.e., $p \geq 0.05$ or trivial $\hat{A}_{12}$) have no rendered color regardless of the results.

Dataset	f_s	Approach	ADF			EIDIG			NeuronFair			SG			ExpGA			BREAM		
			IDI ratio	EOD	SPD	IDI ratio	EOD	SPD	IDI ratio	EOD	SPD	IDI ratio	EOD	SPD	IDI ratio	EOD	SPD	IDI ratio	EOD	SPD
Adult	Gender	w/o CausalFT	.272(.003)	.117(.003)	.086(.003)	.308(.004)	.103(.004)	.089(.003)	.325(.004)	.124(.003)	.107(.003)	.226(.003)	.058(.002)	.075(.004)	.267(.003)	.062(.004)	.061(.003)	.250(.003)	.077(.003)	.076(.003)
		w/ CausalFT	.284(.003)	.137(.003)	.099(.003)	.301(.003)	.153(.003)	.112(.003)	.361(.004)	.145(.002)	.130(.003)	.248(.003)	.120(.003)	.123(.003)	.249(.003)	.091(.003)	.099(.002)	.266(.003)	.102(.003)	.109(.003)
	Race	w/o CausalFT	.336(.002)	.098(.003)	.098(.003)	.362(.003)	.103(.002)	.105(.003)	.380(.003)	.110(.003)	.112(.003)	.220(.003)	.067(.003)	.068(.003)	.222(.002)	.071(.003)	.058(.002)	.285(.003)	.078(.003)	.069(.002)
		w/ CausalFT	.355(.003)	.105(.003)	.125(.003)	.379(.003)	.122(.003)	.129(.003)	.401(.003)	.143(.002)	.145(.003)	.250(.003)	.103(.003)	.107(.003)	.253(.003)	.077(.003)	.080(.002)	.309(.003)	.080(.002)	.085(.003)
	Age	w/o CausalFT	.306(.002)	.095(.002)	.093(.003)	.314(.003)	.109(.002)	.094(.003)	.336(.003)	.128(.002)	.111(.002)	.231(.003)	.055(.002)	.066(.003)	.269(.003)	.057(.002)	.057(.003)	.270(.003)	.063(.003)	.067(.003)
		w/ CausalFT	.333(.003)	.103(.002)	.109(.003)	.327(.002)	.104(.003)	.092(.003)	.389(.003)	.192(.004)	.112(.003)	.232(.003)	.121(.003)	.092(.003)	.255(.003)	.053(.003)	.069(.003)	.278(.003)	.062(.002)	.071(.003)
Compas	Gender	w/o CausalFT	.378(.003)	.096(.002)	.080(.002)	.346(.002)	.116(.002)	.082(.002)	.396(.003)	.137(.003)	.090(.003)	.198(.003)	.057(.002)	.058(.003)	.253(.002)	.068(.003)	.062(.002)	.308(.003)	.051(.003)	.057(.003)
		w/ CausalFT	.363(.003)	.141(.003)	.122(.002)	.362(.003)	.146(.003)	.114(.003)	.419(.003)	.152(.003)	.140(.003)	.231(.003)	.098(.002)	.093(.003)	.252(.003)	.088(.003)	.097(.003)	.312(.003)	.098(.003)	.118(.003)
	Race	w/o CausalFT	.389(.002)	.098(.003)	.077(.002)	.431(.003)	.116(.003)	.095(.002)	.447(.003)	.131(.003)	.103(.003)	.224(.003)	.048(.003)	.064(.003)	.271(.003)	.054(.002)	.075(.003)	.317(.004)	.063(.003)	.065(.003)
		w/ CausalFT	.394(.003)	.131(.002)	.105(.002)	.437(.003)	.150(.003)	.121(.004)	.446(.003)	.165(.002)	.135(.003)	.249(.003)	.087(.002)	.105(.003)	.270(.003)	.067(.003)	.085(.003)	.348(.003)	.089(.003)	.087(.003)
Law School	Gender	w/o CausalFT	.316(.003)	.110(.004)	.082(.002)	.319(.003)	.135(.003)	.092(.002)	.344(.003)	.142(.003)	.096(.003)	.203(.003)	.052(.003)	.052(.003)	.233(.002)	.057(.003)	.061(.003)	.287(.003)	.059(.003)	.069(.003)
		w/ CausalFT	.329(.003)	.118(.003)	.100(.003)	.334(.003)	.141(.003)	.115(.002)	.379(.002)	.158(.003)	.123(.003)	.224(.003)	.106(.003)	.108(.003)	.236(.003)	.066(.003)	.087(.002)	.291(.003)	.077(.003)	.100(.003)
	Race	w/o CausalFT	.293(.002)	.116(.003)	.079(.003)	.317(.003)	.121(.003)	.096(.003)	.368(.002)	.125(.002)	.092(.002)	.214(.002)	.063(.002)	.049(.003)	.233(.003)	.063(.004)	.063(.002)	.271(.003)	.056(.003)	.058(.003)
		w/ CausalFT	.301(.003)	.114(.003)	.089(.003)	.330(.003)	.126(.003)	.105(.003)	.396(.003)	.139(.003)	.126(.003)	.234(.003)	.099(.003)	.098(.003)	.258(.003)	.073(.003)	.069(.003)	.285(.003)	.082(.003)	.089(.002)
Kdd	Gender	w/o CausalFT	.258(.003)	.096(.003)	.081(.003)	.269(.003)	.105(.003)	.086(.003)	.270(.003)	.126(.002)	.099(.003)	.198(.002)	.057(.002)	.059(.002)	.214(.003)	.064(.003)	.072(.003)	.209(.004)	.062(.003)	.072(.003)
		w/ CausalFT	.290(.003)	.129(.003)	.114(.002)	.284(.002)	.137(.003)	.111(.003)	.304(.003)	.160(.003)	.119(.003)	.222(.003)	.090(.003)	.095(.003)	.221(.002)	.077(.002)	.095(.004)	.261(.003)	.081(.003)	.107(.003)
	Race	w/o CausalFT	.271(.002)	.094(.003)	.095(.003)	.270(.002)	.122(.003)	.107(.002)	.269(.003)	.133(.003)	.104(.003)	.191(.003)	.062(.003)	.050(.003)	.219(.003)	.067(.003)	.058(.003)	.231(.003)	.068(.003)	.065(.003)
		w/ CausalFT	.293(.003)	.104(.003)	.111(.002)	.291(.003)	.146(.003)	.110(.003)	.309(.003)	.145(.003)	.125(.003)	.242(.003)	.086(.003)	.097(.003)	.221(.003)	.098(.003)	.103(.003)	.244(.003)	.089(.003)	.114(.003)
Dutch	Gender	w/o CausalFT	.264(.002)	.102(.002)	.091(.002)	.280(.003)	.118(.003)	.090(.003)	.280(.003)	.147(.003)	.095(.003)	.178(.003)	.076(.003)	.057(.003)	.205(.002)	.076(.003)	.079(.004)	.199(.002)	.065(.003)	.080(.004)
		w/ CausalFT	.295(.004)	.127(.003)	.097(.003)	.291(.002)	.139(.003)	.105(.003)	.323(.003)	.148(.003)	.119(.003)	.225(.003)	.110(.002)	.107(.003)	.226(.003)	.093(.003)	.087(.003)	.256(.003)	.081(.003)	.096(.003)
	Age	w/o CausalFT	.278(.003)	.098(.003)	.098(.003)	.284(.003)	.116(.003)	.106(.002)	.273(.004)	.126(.002)	.105(.003)	.176(.002)	.049(.003)	.064(.003)	.215(.003)	.061(.002)	.076(.003)	.201(.003)	.077(.003)	.082(.003)
		w/ CausalFT	.288(.003)	.109(.003)	.119(.003)	.305(.003)	.138(.002)	.113(.003)	.308(.003)	.158(.003)	.135(.003)	.210(.003)	.103(.002)	.107(.003)	.219(.003)	.090(.003)	.096(.003)	.227(.003)	.134(.003)	.113(.003)
Credit	Gender	w/o CausalFT	.251(.003)	.118(.003)	.094(.002)	.274(.003)	.116(.003)	.101(.003)	.282(.002)	.126(.003)	.107(.003)	.185(.003)	.050(.003)	.064(.003)	.194(.003)	.049(.002)	.070(.003)	.206(.003)	.058(.003)	.087(.003)
		w/ CausalFT	.302(.003)	.107(.003)	.109(.003)	.275(.004)	.116(.003)	.094(.003)	.302(.003)	.146(.003)	.101(.003)	.212(.003)	.080(.003)	.095(.003)	.198(.003)	.056(.003)	.080(.003)	.217(.002)	.066(.003)	.089(.003)
	Marriage	w/o CausalFT	.289(.003)	.110(.002)	.076(.003)	.282(.003)	.139(.003)	.086(.003)	.305(.003)	.159(.003)	.091(.003)	.225(.003)	.061(.004)	.059(.002)	.242(.003)	.072(.003)	.066(.003)	.246(.003)	.054(.002)	.061(.002)
		w/ CausalFT	.292(.003)	.113(.003)	.082(.004)	.324(.003)	.122(.002)	.090(.003)	.312(.003)	.138(.003)	.097(.003)	.247(.003)	.099(.003)	.086(.003)	.245(.003)	.073(.003)	.068(.003)	.260(.003)	.077(.003)	.072(.002)
	Education	w/o CausalFT	.272(.002)	.102(.004)	.099(.002)	.285(.003)	.120(.004)	.125(.003)	.291(.003)	.135(.003)	.116(.003)	.204(.003)	.055(.003)	.052(.002)	.216(.004)	.063(.002)	.060(.003)	.211(.003)	.076(.002)	.065(.003)
		w/ CausalFT	.310(.003)	.128(.003)	.145(.003)	.313(.003)	.128(.003)	.138(.003)	.324(.003)	.143(.003)	.120(.003)	.246(.003)	.105(.003)	.093(.003)	.227(.003)	.100(.003)	.120(.003)	.245(.003)	.105(.003)	.112(.003)
Crime	Race	w/o CausalFT	.259(.003)	.124(.003)	.076(.003)	.270(.002)	.135(.004)	.086(.004)	.266(.002)	.153(.002)	.083(.003)	.190(.003)	.051(.002)	.053(.003)	.208(.003)	.061(.003)	.059(.002)	.212(.003)	.062(.003)	.059(.003)
		w/ CausalFT	.278(.003)	.118(.003)	.090(.003)	.284(.002)	.139(.002)	.094(.002)	.309(.003)	.155(.003)	.108(.002)	.254(.003)	.080(.003)	.102(.003)	.237(.003)	.067(.003)	.083(.003)	.257(.003)	.078(.004)	.099(.002)
	Gender	w/o CausalFT	.274(.003)	.113(.002)	.087(.002)	.311(.003)	.122(.003)	.098(.003)	.287(.002)	.137(.003)	.101(.003)	.218(.003)	.094(.004)	.078(.003)	.237(.003)	.104(.003)	.089(.003)	.236(.003)	.054(.003)	.089(.003)
		w/ CausalFT	.315(.003)	.120(.003)	.108(.004)	.318(.002)	.137(.002)	.116(.002)	.332(.002)	.171(.003)	.134(.003)	.249(.003)	.103(.003)	.102(.003)	.256(.002)	.109(.002)	.106(.003)	.273(.003)	.077(.003)	.109(.003)
German	Gender	w/o CausalFT	.270(.003)	.119(.003)	.113(.003)	.313(.003)	.125(.002)	.111(.003)	.316(.003)	.141(.002)	.123(.003)	.190(.002)	.072(.003)	.057(.004)	.204(.002)	.080(.003)	.060(.003)	.216(.003)	.104(.002)	.069(.004)
		w/ CausalFT	.321(.004)	.118(.002)	.105(.003)	.303(.003)	.141(.003)	.145(.003)	.331(.003)	.154(.003)	.149(.002)	.229(.004)	.088(.003)	.074(.002)	.225(.003)	.090(.003)	.078(.002)	.266(.002)	.114(.004)	.092(.003)
	Age	w/o CausalFT	.259(.003)	.094(.002)	.075(.003)	.266(.003)	.111(.003)	.080(.003)	.283(.003)	.112(.003)	.089(.003)	.207(.003)	.081(.003)	.071(.003)	.229(.003)	.086(.002)	.065(.003)	.225(.003)	.084(.002)	.070(.004)
		w/ CausalFT	.260(.003)	.131(.004)	.101(.003)	.283(.003)	.146(.003)	.122(.003)	.310(.003)	.143(.003)	.127(.003)	.264(.003)	.116(.003)	.116(.003)	.240(.003)	.109(.003)	.106(.003)	.257(.003)	.091(.003)	.103(.003)

(3) Randomly make 70%/30% training and testing data split and use the training data to train three different AI models as used in existing work [14, 79].
(4) For an approach that can rank non-sensitive features, such as CausalFT, perform analysis in the training data and select a representative non-sensitive feature.
(5) Pick a test generator and run it based on the 30% testing data to generate 10,000 unique test samples, with and without handling the relationship between the selected non-sensitive feature and the sensitive counterpart.
(6) Evaluate the results using IDI ratio, EOD, and SPD.
(7) Repeat from 5) until all generators have been examined.
(8) Repeat from 4) until CausalFT and other state-of-the-art approaches have been evaluated.
(9) Repeat from 3) via bootstrapping (with replacement) for 10 runs.
(10) Repeat from 2) until all possible sensitive features in the dataset have served as the concerned feature once.
(11) Repeat from 1) until all datasets have been used.

4.6 Statistical Validation

We use the recommended non-parametric U-Test [65] with $a = 0.05$ to verify the significance of pairwise comparisons over 10 runs on each metric [10]. We additionally use $\hat{A}_{12}$ [64] to examine the effect size. According to Vargha and Delaney [64], $\hat{A}_{12} \geq 0.56$ (or $\hat{A}_{12} \leq 0.44$) indicates a non-trivial effect size. In this work, we say the difference is statistically significant only when $\hat{A}_{12} \geq 0.56$ (or $\hat{A}_{12} \leq 0.44$) and $p < 0.05$; otherwise the deviation is trivial.

Table 3. Comparing different generators with and without CausalFT over all cases using the mean (standard deviation) of individual/group fairness metrics under 10 runs when testing DNN_6. All formate are the same as Table 2.

Dataset	f_s	Approach	ADF			EIDIG			NeuronFair			SG			ExpGA			BREAM		
			IDI ratio	EOD	SPD	IDI ratio	EOD	SPD	IDI ratio	EOD	SPD	IDI ratio	EOD	SPD	IDI ratio	EOD	SPD	IDI ratio	EOD	SPD
Adult	*Gender*	w/o CausalFT	.291(.003)	.124(.003)	.086(.003)	.333(.004)	.114(.004)	.097(.003)	.351(.004)	.138(.003)	.115(.003)	.241(.003)	.064(.002)	.082(.010)	.266(.003)	.068(.004)	.064(.003)	.270(.003)	.077(.003)	.076(.003)
		w/ CausalFT	.301(.003)	.142(.003)	.103(.003)	.316(.003)	.169(.003)	.120(.003)	.365(.004)	.154(.002)	.137(.003)	.262(.003)	.127(.003)	.116(.003)	.274(.003)	.092(.003)	.104(.002)	.284(.003)	.109(.003)	.121(.003)
	Race	w/o CausalFT	.361(.002)	.105(.003)	.102(.003)	.382(.003)	.110(.002)	.110(.003)	.413(.003)	.120(.003)	.124(.003)	.238(.003)	.064(.003)	.071(.003)	.240(.002)	.067(.003)	.061(.002)	.299(.003)	.081(.003)	.076(.002)
		w/ CausalFT	.382(.003)	.113(.003)	.132(.003)	.396(.003)	.131(.003)	.145(.003)	.429(.003)	.154(.002)	.155(.003)	.249(.003)	.102(.003)	.116(.003)	.251(.003)	.079(.003)	.079(.001)	.316(.003)	.086(.002)	.097(.003)
	Age	w/o CausalFT	.327(.002)	.103(.002)	.098(.003)	.343(.003)	.119(.002)	.106(.003)	.365(.003)	.139(.002)	.116(.002)	.250(.003)	.061(.002)	.071(.003)	.274(.003)	.063(.002)	.065(.003)	.296(.003)	.071(.003)	.071(.003)
		w/ CausalFT	.353(.003)	.111(.002)	.117(.003)	.357(.002)	.113(.004)	.104(.003)	.388(.003)	.209(.004)	.122(.003)	.251(.003)	.122(.003)	.098(.003)	.274(.003)	.056(.003)	.069(.003)	.308(.003)	.063(.002)	.074(.003)
Compas	*Gender*	w/o CausalFT	.364(.003)	.102(.002)	.085(.002)	.381(.002)	.122(.002)	.092(.002)	.425(.003)	.136(.003)	.098(.003)	.210(.003)	.061(.002)	.054(.003)	.269(.002)	.067(.003)	.068(.002)	.335(.003)	.056(.003)	.064(.003)
		w/ CausalFT	.386(.003)	.144(.003)	.130(.002)	.394(.003)	.157(.003)	.124(.003)	.419(.003)	.163(.003)	.153(.003)	.247(.003)	.103(.002)	.094(.003)	.272(.003)	.092(.003)	.105(.003)	.335(.003)	.110(.003)	.126(.003)
	Race	w/o CausalFT	.418(.002)	.104(.003)	.084(.002)	.426(.003)	.119(.003)	.091(.002)	.472(.003)	.141(.003)	.107(.003)	.240(.003)	.050(.003)	.066(.003)	.286(.003)	.056(.002)	.083(.003)	.343(.004)	.068(.003)	.077(.003)
		w/ CausalFT	.426(.003)	.138(.002)	.113(.002)	.435(.003)	.158(.003)	.130(.004)	.482(.003)	.175(.002)	.140(.003)	.272(.003)	.094(.002)	.112(.003)	.297(.003)	.070(.003)	.087(.003)	.363(.003)	.083(.003)	.095(.003)
Law School	*Gender*	w/o CausalFT	.322(.003)	.122(.004)	.083(.002)	.348(.003)	.132(.003)	.092(.002)	.371(.003)	.154(.003)	.102(.003)	.220(.003)	.055(.003)	.059(.003)	.250(.002)	.061(.003)	.068(.003)	.311(.003)	.059(.003)	.073(.003)
		w/ CausalFT	.334(.003)	.128(.003)	.107(.003)	.356(.003)	.150(.003)	.125(.002)	.398(.002)	.166(.003)	.133(.003)	.239(.003)	.107(.003)	.121(.003)	.254(.003)	.067(.003)	.095(.002)	.313(.003)	.082(.003)	.107(.003)
	Race	w/o CausalFT	.316(.002)	.117(.003)	.084(.003)	.342(.003)	.127(.003)	.094(.003)	.386(.002)	.139(.002)	.100(.002)	.224(.002)	.065(.002)	.054(.003)	.256(.003)	.071(.004)	.065(.002)	.291(.003)	.061(.003)	.063(.003)
		w/ CausalFT	.322(.003)	.114(.003)	.092(.003)	.333(.003)	.138(.003)	.113(.003)	.385(.003)	.149(.003)	.128(.003)	.251(.003)	.104(.003)	.111(.003)	.273(.003)	.079(.003)	.079(.003)	.308(.003)	.079(.003)	.096(.002)
Kdd	*Gender*	w/o CausalFT	.274(.003)	.106(.003)	.085(.003)	.288(.003)	.118(.003)	.092(.003)	.291(.003)	.133(.002)	.107(.003)	.208(.002)	.064(.002)	.061(.002)	.228(.003)	.071(.003)	.076(.003)	.226(.004)	.073(.003)	.082(.003)
		w/ CausalFT	.308(.003)	.134(.003)	.116(.002)	.309(.002)	.148(.003)	.118(.003)	.324(.003)	.170(.003)	.124(.003)	.242(.003)	.097(.003)	.103(.003)	.246(.002)	.079(.002)	.102(.004)	.260(.003)	.086(.003)	.119(.003)
	Race	w/o CausalFT	.276(.002)	.108(.003)	.094(.003)	.292(.002)	.120(.003)	.105(.002)	.295(.003)	.137(.003)	.104(.003)	.208(.003)	.062(.003)	.056(.003)	.229(.003)	.068(.003)	.065(.003)	.227(.003)	.071(.003)	.071(.003)
		w/ CausalFT	.311(.003)	.116(.003)	.121(.002)	.317(.003)	.142(.003)	.119(.003)	.332(.003)	.152(.003)	.143(.003)	.238(.003)	.097(.003)	.104(.003)	.245(.003)	.104(.003)	.111(.003)	.262(.003)	.092(.003)	.120(.003)
Dutch	*Gender*	w/o CausalFT	.282(.002)	.111(.002)	.094(.002)	.300(.003)	.127(.003)	.098(.003)	.302(.003)	.145(.003)	.108(.003)	.194(.003)	.076(.003)	.060(.003)	.215(.002)	.080(.003)	.087(.004)	.217(.002)	.072(.003)	.081(.004)
		w/ CausalFT	.319(.004)	.134(.003)	.100(.003)	.322(.002)	.149(.003)	.112(.003)	.344(.003)	.162(.003)	.127(.003)	.235(.003)	.105(.002)	.112(.003)	.238(.003)	.095(.003)	.095(.003)	.252(.003)	.087(.003)	.102(.003)
	Age	w/o CausalFT	.279(.003)	.108(.003)	.106(.003)	.300(.003)	.123(.003)	.110(.002)	.296(.004)	.136(.002)	.118(.003)	.193(.002)	.055(.003)	.069(.003)	.211(.003)	.064(.002)	.081(.003)	.215(.003)	.084(.003)	.090(.003)
		w/ CausalFT	.313(.003)	.123(.003)	.131(.003)	.321(.003)	.146(.002)	.124(.003)	.334(.003)	.164(.003)	.144(.003)	.225(.003)	.102(.002)	.116(.003)	.233(.003)	.094(.003)	.103(.003)	.247(.003)	.142(.003)	.122(.003)
Credit	*Gender*	w/o CausalFT	.275(.003)	.117(.003)	.101(.002)	.293(.003)	.129(.003)	.109(.003)	.299(.002)	.140(.003)	.113(.003)	.195(.003)	.051(.003)	.069(.003)	.214(.003)	.055(.002)	.077(.003)	.217(.003)	.063(.003)	.093(.003)
		w/ CausalFT	.292(.003)	.114(.003)	.108(.003)	.299(.004)	.124(.003)	.104(.003)	.316(.003)	.157(.003)	.111(.003)	.219(.003)	.086(.003)	.101(.003)	.215(.003)	.054(.003)	.088(.003)	.237(.002)	.069(.003)	.093(.003)
	Marriage	w/o CausalFT	.286(.003)	.121(.002)	.081(.003)	.303(.003)	.142(.003)	.092(.003)	.307(.003)	.153(.003)	.098(.003)	.245(.003)	.070(.004)	.061(.002)	.263(.003)	.076(.003)	.073(.003)	.267(.003)	.058(.002)	.070(.002)
		w/ CausalFT	.308(.003)	.118(.003)	.089(.004)	.317(.003)	.135(.002)	.093(.003)	.324(.003)	.149(.003)	.102(.003)	.263(.003)	.101(.003)	.092(.003)	.261(.003)	.073(.003)	.067(.003)	.284(.003)	.079(.003)	.077(.002)
	Education	w/o CausalFT	.292(.002)	.108(.004)	.110(.002)	.310(.003)	.127(.004)	.136(.003)	.313(.003)	.139(.003)	.123(.003)	.210(.003)	.058(.003)	.056(.002)	.232(.004)	.066(.002)	.065(.003)	.228(.003)	.081(.002)	.070(.003)
		w/ CausalFT	.324(.003)	.136(.003)	.154(.003)	.327(.003)	.135(.003)	.142(.003)	.343(.003)	.156(.003)	.119(.003)	.242(.003)	.102(.003)	.097(.003)	.245(.003)	.110(.003)	.127(.003)	.263(.003)	.113(.003)	.121(.003)
Crime	*Race*	w/o CausalFT	.275(.003)	.132(.003)	.081(.003)	.291(.002)	.144(.004)	.089(.004)	.292(.002)	.167(.002)	.091(.003)	.202(.003)	.053(.002)	.054(.003)	.217(.003)	.058(.003)	.066(.002)	.225(.003)	.065(.003)	.064(.003)
		w/ CausalFT	.307(.003)	.126(.003)	.100(.003)	.309(.002)	.150(.002)	.103(.002)	.331(.003)	.169(.003)	.115(.002)	.246(.003)	.082(.003)	.102(.003)	.248(.003)	.071(.003)	.087(.003)	.269(.003)	.082(.004)	.097(.002)
	Gender	w/o CausalFT	.295(.003)	.117(.002)	.091(.002)	.304(.003)	.129(.003)	.102(.003)	.311(.002)	.146(.003)	.107(.003)	.232(.003)	.103(.004)	.075(.003)	.253(.003)	.113(.003)	.091(.003)	.255(.003)	.059(.003)	.095(.003)
		w/ CausalFT	.332(.003)	.131(.003)	.114(.004)	.330(.002)	.145(.002)	.126(.002)	.352(.002)	.168(.003)	.136(.003)	.271(.003)	.114(.003)	.113(.003)	.278(.002)	.117(.002)	.113(.003)	.294(.003)	.081(.003)	.119(.003)
German	*Gender*	w/o CausalFT	.296(.003)	.127(.003)	.111(.003)	.310(.003)	.136(.002)	.120(.003)	.318(.003)	.150(.002)	.131(.003)	.206(.002)	.076(.003)	.060(.004)	.221(.002)	.085(.003)	.068(.003)	.223(.003)	.112(.002)	.074(.004)
		w/ CausalFT	.317(.004)	.129(.002)	.112(.003)	.324(.003)	.139(.003)	.145(.003)	.333(.003)	.150(.003)	.147(.002)	.241(.004)	.097(.003)	.082(.002)	.246(.003)	.094(.003)	.084(.002)	.264(.002)	.121(.004)	.095(.003)
	Age	w/o CausalFT	.276(.003)	.105(.002)	.081(.003)	.288(.003)	.114(.003)	.087(.003)	.296(.003)	.124(.003)	.096(.003)	.223(.003)	.085(.003)	.080(.003)	.245(.003)	.096(.002)	.069(.003)	.242(.003)	.087(.002)	.075(.004)
		w/ CausalFT	.284(.003)	.142(.004)	.113(.003)	.311(.003)	.154(.003)	.127(.003)	.329(.003)	.156(.003)	.134(.003)	.258(.003)	.124(.003)	.129(.003)	.259(.003)	.125(.003)	.111(.003)	.280(.003)	.103(.003)	.114(.003)

Table 4. Comparing different generators with and without CausalFT over all cases using the mean(standard deviation) of individual/group fairness metrics under 10 runs when testing LR. All formate are the same as Table 2.

Dataset	f_s	Approach	ADF			EIDIG			NeuronFair			SG			ExpGA			BREAM		
			IDI ratio	EOD	SPD	IDI ratio	EOD	SPD	IDI ratio	EOD	SPD	IDI ratio	EOD	SPD	IDI ratio	EOD	SPD	IDI ratio	EOD	SPD
Adult	*Gender*	w/o CausalFT	.175(.003)	.085(.003)	.198(.003)	.182(.003)	.105(.002)	.187(.001)	.185(.002)	.091(.002)	.146(.002)	.150(.004)	.122(.004)	.174(.006)	.181(.003)	.144(.001)	.099(.004)	.169(.003)	.178(.003)	.083(.003)
		w/ CausalFT	.180(.003)	.080(.003)	.206(.003)	.189(.003)	.112(.002)	.193(.001)	.191(.002)	.096(.002)	.152(.002)	.156(.004)	.127(.004)	.180(.006)	.188(.003)	.150(.001)	.105(.004)	.175(.003)	.185(.003)	.088(.003)
	Race	w/o CausalFT	.172(.003)	.108(.005)	.208(.004)	.162(.003)	.117(.004)	.074(.003)	.179(.002)	.201(.003)	.166(.002)	.167(.002)	.121(.004)	.172(.006)	.175(.002)	.173(.003)	.122(.003)	.163(.003)	.140(.002)	.080(.003)
		w/ CausalFT	.178(.003)	.114(.005)	.217(.004)	.169(.003)	.123(.004)	.073(.003)	.176(.002)	.210(.003)	.173(.002)	.174(.002)	.127(.004)	.180(.006)	.172(.002)	.180(.003)	.128(.003)	.170(.003)	.146(.002)	.079(.003)
	Age	w/o CausalFT	.155(.002)	.147(.003)	.224(.003)	.164(.002)	.135(.003)	.219(.003)	.180(.002)	.138(.003)	.228(.003)	.144(.002)	.152(.003)	.210(.003)	.146(.002)	.134(.003)	.223(.003)	.151(.002)	.138(.003)	.227(.003)
		w/ CausalFT	.169(.002)	.159(.003)	.242(.003)	.178(.002)	.146(.003)	.236(.003)	.196(.002)	.149(.003)	.246(.003)	.143(.002)	.164(.003)	.227(.003)	.159(.002)	.145(.003)	.240(.003)	.164(.002)	.149(.003)	.245(.003)
Compas	*Gender*	w/o CausalFT	.164(.002)	.189(.004)	.108(.003)	.155(.002)	.075(.002)	.224(.002)	.176(.002)	.181(.002)	.259(.002)	.145(.003)	.117(.004)	.166(.004)	.166(.003)	.136(.004)	.111(.004)	.165(.003)	.179(.001)	.176(.002)
		w/ CausalFT	.179(.002)	.206(.004)	.117(.003)	.169(.002)	.074(.002)	.243(.002)	.191(.002)	.197(.002)	.281(.002)	.158(.003)	.127(.004)	.180(.004)	.180(.003)	.148(.004)	.121(.004)	.179(.003)	.195(.001)	.191(.002)
	Race	w/o CausalFT	.159(.003)	.098(.002)	.081(.004)	.166(.003)	.148(.004)	.100(.002)	.179(.003)	.136(.004)	.117(.004)	.140(.003)	.117(.004)	.166(.003)	.147(.003)	.090(.003)	.198(.002)	.156(.002)	.090(.002)	.149(.001)
		w/ CausalFT	.173(.003)	.107(.002)	.080(.004)	.180(.003)	.161(.004)	.109(.002)	.195(.003)	.148(.004)	.127(.004)	.152(.003)	.127(.004)	.180(.003)	.160(.003)	.089(.003)	.215(.002)	.170(.002)	.098(.002)	.162(.001)
Law School	*Gender*	w/o CausalFT	.149(.003)	.075(.004)	.107(.004)	.163(.003)	.101(.004)	.078(.004)	.164(.003)	.144(.004)	.179(.002)	.148(.002)	.136(.002)	.219(.003)	.143(.003)	.165(.002)	.220(.001)	.152(.003)	.093(.005)	.160(.004)
		w/ CausalFT	.162(.003)	.074(.004)	.117(.004)	.177(.003)	.110(.004)	.085(.004)	.178(.003)	.157(.004)	.195(.002)	.161(.002)	.148(.002)	.238(.003)	.155(.003)	.179(.002)	.239(.001)	.165(.003)	.092(.005)	.174(.004)
	Race	w/o CausalFT	.158(.002)	.120(.004)	.107(.004)	.160(.002)	.150(.004)	.210(.004)	.163(.003)	.140(.004)	.165(.003)	.143(.003)	.181(.002)	.172(.002)	.161(.003)	.194(.001)	.190(.002)	.152(.002)	.184(.004)	.215(.005)
		w/ CausalFT	.175(.002)	.133(.004)	.119(.004)	.178(.002)	.167(.004)	.233(.004)	.181(.003)	.139(.004)	.183(.003)	.159(.003)	.201(.002)	.191(.002)	.179(.003)	.215(.001)	.211(.002)	.169(.002)	.204(.004)	.239(.005)
Kdd	*Gender*	w/o CausalFT	.152(.003)	.117(.004)	.070(.001)	.158(.003)	.159(.004)	.071(.001)	.162(.003)	.069(.002)	.203(.004)	.153(.003)	.105(.002)	.169(.003)	.152(.003)	.089(.004)	.203(.004)	.154(.004)	.097(.004)	.139(.001)
		w/ CausalFT	.168(.003)	.130(.004)	.069(.001)	.175(.003)	.177(.004)	.078(.001)	.180(.003)	.076(.002)	.225(.004)	.170(.003)	.116(.002)	.187(.003)	.169(.003)	.099(.004)	.225(.004)	.171(.004)	.096(.004)	.154(.001)
	Race	w/o CausalFT	.160(.002)	.186(.002)	.247(.003)	.154(.003)	.111(.003)	.144(.004)	.159(.002)	.084(.002)	.212(.003)	.149(.002)	.091(.002)	.184(.003)	.154(.003)	.087(.002)	.206(.004)	.152(.002)	.090(.002)	.211(.003)
		w/ CausalFT	.178(.002)	.207(.002)	.275(.003)	.171(.003)	.110(.003)	.160(.004)	.177(.002)	.093(.002)	.236(.003)	.166(.002)	.101(.002)	.205(.003)	.171(.003)	.097(.002)	.229(.004)	.169(.002)	.100(.002)	.235(.003)
Dutch	*Gender*	w/o CausalFT	.153(.002)	.096(.004)	.232(.003)	.144(.002)	.077(.001)	.170(.003)	.154(.003)	.124(.003)	.209(.003)	.144(.002)	.149(.004)	.205(.003)	.161(.003)	.090(.004)	.176(.002)	.156(.002)	.075(.002)	.092(.004)
		w/ CausalFT	.174(.002)	.109(.004)	.258(.003)	.164(.002)	.076(.001)	.189(.003)	.175(.003)	.141(.003)	.232(.003)	.164(.002)	.169(.004)	.228(.003)	.183(.003)	.102(.004)	.196(.002)	.177(.002)	.074(.002)	.102(.004)
	Age	w/o CausalFT	.144(.003)	.075(.001)	.082(.003)	.149(.003)	.155(.002)	.161(.004)	.171(.002)	.168(.004)	.072(.002)	.154(.003)	.192(.003)	.214(.002)	.163(.003)	.183(.003)	.223(.002)	.165(.003)	.100(.005)	.094(.004)
		w/ CausalFT	.164(.003)	.074(.001)	.093(.003)	.169(.003)	.176(.002)	.183(.004)	.194(.002)	.191(.004)	.082(.002)	.175(.003)	.218(.003)	.243(.002)	.185(.003)	.208(.003)	.253(.002)	.187(.003)	.114(.005)	.107(.004)
Credit	*Gender*	w/o CausalFT	.144(.003)	.075(.001)	.082(.003)	.149(.003)	.155(.002)	.161(.004)	.171(.002)	.168(.004)	.072(.002)	.154(.003)	.192(.003)	.214(.002)	.163(.003)	.183(.003)	.223(.002)	.165(.003)	.100(.005)	.094(.004)
		w/ CausalFT	.171(.002)	.074(.001)	.120(.003)	.175(.002)	.093(.002)	.067(.004)	.166(.002)	.077(.001)	.270(.004)	.160(.002)	.151(.003)	.148(.002)	.182(.003)	.162(.003)	.090(.003)	.170(.002)	.131(.003)	.188(.002)
	Marriage	w/o CausalFT	.144(.001)	.116(.004)	.143(.001)	.165(.003)	.092(.003)	.106(.002)	.168(.003)	.106(.003)	.150(.003)	.148(.002)	.193(.001)	.232(.002)	.133(.003)	.177(.003)	.212(.002)	.149(.001)	.199(.002)	.228(.002)
		w/ CausalFT	.164(.001)	.132(.004)	.162(.001)	.188(.003)	.091(.003)	.120(.002)	.191(.003)	.120(.003)	.170(.003)	.168(.002)	.219(.001)	.264(.002)	.151(.003)	.201(.003)	.241(.002)	.169(.001)	.226(.002)	.259(.002)
	Education	w/o CausalFT	.177(.002)	.148(.003)	.225(.003)	.185(.002)	.143(.003)	.231(.003)	.199(.002)	.164(.003)	.230(.003)	.153(.002)	.159(.003)	.213(.003)	.166(.002)	.166(.003)	.223(.003)	.164(.002)	.139(.003)	.204(.003)
		w/ CausalFT	.201(.002)	.168(.003)	.256(.003)	.210(.002)	.162(.003)	.263(.003)	.226(.002)	.163(.003)	.262(.003)	.174(.002)	.181(.003)	.242(.003)	.189(.002)	.189(.003)	.253(.003)	.186(.002)	.158(.003)	.232(.003)
Crime	*Race*	w/o CausalFT	.148(.003)	.076(.002)	.143(.002)	.150(.003)	.063(.005)	.189(.004)	.152(.003)	.193(.005)	.124(.003)	.138(.002)	.126(.002)	.094(.003)	.143(.002)	.156(.002)	.112(.003)	.153(.003)	.140(.004)	.057(.002)
		w/ CausalFT	.147(.003)	.086(.002)	.162(.002)	.174(.003)	.071(.005)	.215(.004)	.173(.003)	.219(.005)	.141(.003)	.157(.002)	.143(.002)	.107(.003)	.162(.002)	.177(.002)	.127(.003)	.172(.003)	.159(.004)	.056(.002)
	Gender	w/o CausalFT	.161(.002)	.142(.003)	.118(.002)	.178(.003)	.103(.003)	.227(.002)	.187(.002)	.062(.001)	.199(.002)	.157(.002)	.094(.002)	.143(.002)	.167(.002)	.104(.002)	.151(.002)	.177(.003)	.115(.002)	.283(.002)
		w/ CausalFT	.168(.002)	.149(.003)	.125(.002)	.186(.003)	.102(.003)	.234(.002)	.183(.002)	.069(.001)	.206(.002)	.164(.002)	.101(.002)	.150(.002)	.174(.002)	.111(.002)	.158(.002)	.182(.003)	.122(.002)	.290(.002)
German	*Gender*	w/o CausalFT	.157(.002)	.109(.003)	.074(.004)	.155(.002)	.150(.003)	.239(.002)	.158(.003)	.095(.003)	.179(.004)	.150(.003)	.112(.004)	.158(.003)	.150(.003)	.121(.003)	.193(.002)	.149(.002)	.186(.005)	.228(.002)
		w/ CausalFT	.178(.002)	.108(.003)	.084(.004)	.176(.002)	.171(.003)	.272(.002)	.177(.003)	.108(.003)	.203(.004)	.171(.003)	.127(.004)	.180(.003)	.170(.003)	.138(.003)	.219(.002)	.169(.002)	.212(.005)	.259(.002)
	Age	w/o CausalFT	.157(.002)	.085(.002)	.133(.004)	.153(.003)	.130(.005)	.107(.003)	.167(.002)	.172(.004)	.133(.005)	.142(.002)	.112(.004)	.158(.006)	.154(.002)	.148(.002)	.094(.002)	.147(.003)	.152(.003)	.132(.004)
		w/ CausalFT	.178(.002)	.096(.002)	.132(.004)	.174(.003)	.148(.005)	.122(.003)	.190(.002)	.195(.004)	.151(.005)	.161(.002)	.127(.004)	.180(.006)	.175(.002)	.168(.002)	.107(.002)	.167(.003)	.173(.003)	.131(.004)

5 Results

5.1 Improvement over State-of-the-art Generators

5.1.1 Method. In $\mathbf{RQ}_1$, to evaluate the generator-agnostic nature of CausalFT and its benefits to the state-of-the-art generators for fairness testing, we pair CausalFT with all six generators studied

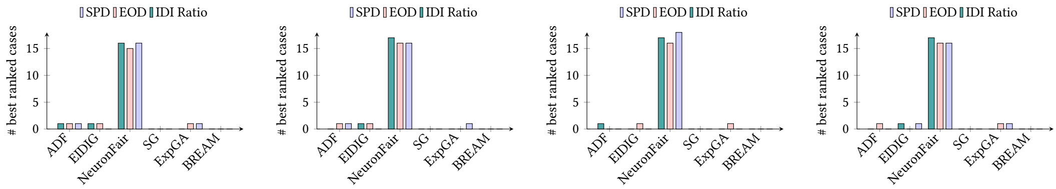
Table 5. Comparing different generators with and without CausalFT over all cases using the mean (standard deviation) of individual/group fairness metrics under 10 runs when testing RF. All formate are the same as Table 2.

Dataset	f_s	Approach	ADF			EIDIG			NeuronFair			SG			ExpGA			BREAM		
			IDI ratio	EOD	SPD	IDI ratio	EOD	SPD	IDI ratio	EOD	SPD	IDI ratio	EOD	SPD	IDI ratio	EOD	SPD	IDI ratio	EOD	SPD
Adult	*Gender*	w/o CausalFT	.155 (.003)	.075 (.004)	.131 (.003)	.167 (.003)	.146 (.004)	.157 (.003)	.179 (.003)	.153 (.005)	.181 (.004)	.139 (.003)	.093 (.003)	.081 (.003)	.153 (.003)	.123 (.004)	.102 (.003)	.173 (.003)	.160 (.005)	.188 (.004)
		w/ CausalFT	.165 (.003)	.079 (.004)	.129 (.003)	.171 (.003)	.145 (.004)	.167 (.003)	.186 (.003)	.159 (.005)	.189 (.004)	.159 (.003)	.106 (.003)	.080 (.003)	.171 (.003)	.131 (.004)	.110 (.003)	.197 (.003)	.172 (.005)	.201 (.004)
	Race	w/o CausalFT	.150 (.003)	.128 (.003)	.090 (.003)	.169 (.003)	.146 (.004)	.101 (.003)	.181 (.003)	.188 (.005)	.174 (.004)	.140 (.003)	.109 (.003)	.082 (.003)	.163 (.003)	.135 (.004)	.108 (.003)	.178 (.003)	.204 (.005)	.195 (.004)
		w/ CausalFT	.163 (.003)	.130 (.003)	.089 (.003)	.176 (.003)	.157 (.004)	.106 (.003)	.189 (.003)	.199 (.004)	.186 (.004)	.152 (.003)	.123 (.003)	.088 (.003)	.174 (.003)	.147 (.004)	.116 (.003)	.194 (.003)	.221 (.005)	.208 (.004)
	Age	w/o CausalFT	.141 (.003)	.111 (.003)	.122 (.003)	.157 (.003)	.126 (.004)	.139 (.003)	.163 (.003)	.144 (.005)	.156 (.004)	.138 (.003)	.093 (.003)	.088 (.003)	.150 (.003)	.116 (.004)	.100 (.003)	.171 (.003)	.132 (.005)	.112 (.004)
		w/ CausalFT	.140 (.003)	.114 (.003)	.127 (.003)	.162 (.003)	.130 (.004)	.143 (.003)	.181 (.003)	.155 (.005)	.173 (.004)	.151 (.003)	.102 (.003)	.091 (.003)	.166 (.003)	.127 (.004)	.108 (.003)	.185 (.003)	.142 (.005)	.125 (.004)
Compas	*Gender*	w/o CausalFT	.143 (.002)	.076 (.003)	.105 (.002)	.154 (.002)	.123 (.004)	.122 (.003)	.175 (.003)	.130 (.005)	.191 (.004)	.131 (.002)	.073 (.003)	.085 (.002)	.147 (.002)	.124 (.004)	.111 (.003)	.164 (.003)	.172 (.005)	.271 (.004)
		w/ CausalFT	.143 (.002)	.077 (.003)	.105 (.002)	.152 (.002)	.122 (.004)	.121 (.003)	.179 (.003)	.135 (.005)	.207 (.004)	.142 (.002)	.079 (.003)	.091 (.002)	.160 (.002)	.131 (.004)	.118 (.003)	.174 (.003)	.186 (.005)	.289 (.004)
	Race	w/o CausalFT	.133 (.002)	.094 (.003)	.077 (.002)	.149 (.002)	.116 (.004)	.091 (.003)	.162 (.002)	.141 (.005)	.109 (.004)	.128 (.002)	.088 (.003)	.073 (.002)	.143 (.002)	.108 (.004)	.100 (.003)	.163 (.003)	.125 (.005)	.122 (.004)
		w/ CausalFT	.148 (.002)	.093 (.003)	.085 (.002)	.156 (.002)	.128 (.004)	.100 (.003)	.169 (.002)	.154 (.005)	.123 (.004)	.142 (.002)	.096 (.003)	.082 (.002)	.160 (.002)	.119 (.004)	.112 (.003)	.180 (.003)	.139 (.005)	.136 (.004)
Law School	*Gender*	w/o CausalFT	.148 (.003)	.105 (.004)	.073 (.002)	.158 (.003)	.121 (.003)	.104 (.003)	.170 (.003)	.147 (.004)	.198 (.004)	.137 (.003)	.107 (.003)	.075 (.002)	.156 (.003)	.142 (.004)	.111 (.003)	.166 (.003)	.193 (.005)	.244 (.004)
		w/ CausalFT	.153 (.003)	.108 (.004)	.072 (.002)	.163 (.003)	.125 (.003)	.109 (.003)	.175 (.003)	.152 (.004)	.206 (.004)	.141 (.003)	.110 (.003)	.077 (.002)	.161 (.003)	.146 (.004)	.116 (.003)	.171 (.003)	.199 (.005)	.251 (.004)
	Race	w/o CausalFT	.144 (.003)	.109 (.003)	.082 (.003)	.153 (.003)	.123 (.004)	.092 (.003)	.166 (.003)	.138 (.005)	.106 (.004)	.128 (.003)	.096 (.003)	.071 (.003)	.148 (.003)	.112 (.004)	.085 (.003)	.158 (.003)	.129 (.005)	.100 (.004)
		w/ CausalFT	.148 (.003)	.112 (.003)	.085 (.003)	.158 (.003)	.122 (.004)	.097 (.003)	.171 (.003)	.143 (.005)	.110 (.004)	.132 (.003)	.099 (.003)	.073 (.003)	.153 (.003)	.117 (.004)	.088 (.003)	.163 (.003)	.134 (.005)	.103 (.004)
Kdd	*Gender*	w/o CausalFT	.144 (.003)	.097 (.003)	.126 (.003)	.163 (.003)	.129 (.004)	.143 (.004)	.165 (.003)	.183 (.005)	.228 (.004)	.134 (.003)	.103 (.003)	.103 (.003)	.152 (.003)	.139 (.004)	.131 (.003)	.172 (.003)	.242 (.005)	.167 (.004)
		w/ CausalFT	.148 (.003)	.099 (.003)	.130 (.003)	.162 (.003)	.133 (.004)	.148 (.004)	.170 (.003)	.189 (.005)	.235 (.004)	.138 (.003)	.106 (.003)	.106 (.003)	.157 (.003)	.143 (.004)	.135 (.003)	.177 (.003)	.249 (.005)	.172 (.004)
	Race	w/o CausalFT	.147 (.003)	.142 (.003)	.166 (.003)	.159 (.003)	.151 (.004)	.174 (.003)	.168 (.003)	.192 (.005)	.243 (.004)	.141 (.003)	.146 (.003)	.157 (.003)	.156 (.003)	.179 (.004)	.196 (.003)	.176 (.003)	.247 (.005)	.203 (.004)
		w/ CausalFT	.152 (.003)	.146 (.003)	.171 (.003)	.164 (.003)	.155 (.004)	.179 (.003)	.173 (.003)	.198 (.005)	.250 (.004)	.145 (.003)	.151 (.003)	.162 (.003)	.161 (.003)	.184 (.004)	.202 (.003)	.181 (.003)	.253 (.005)	.209 (.004)
Dutch	*Gender*	w/o CausalFT	.159 (.003)	.074 (.003)	.126 (.003)	.176 (.003)	.097 (.004)	.128 (.004)	.170 (.003)	.141 (.005)	.237 (.004)	.157 (.003)	.089 (.003)	.115 (.003)	.166 (.003)	.122 (.004)	.184 (.003)	.170 (.003)	.261 (.005)	.214 (.004)
		w/ CausalFT	.164 (.003)	.076 (.003)	.130 (.003)	.181 (.003)	.100 (.004)	.132 (.004)	.175 (.003)	.146 (.005)	.244 (.004)	.162 (.003)	.092 (.003)	.119 (.003)	.171 (.003)	.126 (.004)	.190 (.003)	.175 (.003)	.269 (.005)	.221 (.004)
	Age	w/o CausalFT	.157 (.003)	.125 (.003)	.109 (.003)	.176 (.003)	.140 (.004)	.179 (.003)	.174 (.003)	.203 (.005)	.252 (.004)	.154 (.003)	.142 (.003)	.156 (.003)	.169 (.003)	.193 (.004)	.184 (.003)	.175 (.003)	.222 (.005)	.198 (.004)
		w/ CausalFT	.162 (.003)	.129 (.003)	.113 (.003)	.181 (.003)	.144 (.004)	.185 (.003)	.179 (.003)	.202 (.005)	.260 (.004)	.159 (.003)	.147 (.003)	.161 (.003)	.174 (.003)	.199 (.004)	.190 (.003)	.180 (.003)	.229 (.005)	.204 (.004)
Credit	*Gender*	w/o CausalFT	.154 (.003)	.074 (.003)	.117 (.003)	.162 (.003)	.130 (.004)	.131 (.003)	.167 (.003)	.163 (.005)	.184 (.004)	.153 (.003)	.071 (.003)	.092 (.003)	.164 (.003)	.179 (.004)	.158 (.003)	.167 (.003)	.219 (.005)	.208 (.004)
		w/ CausalFT	.159 (.003)	.077 (.003)	.121 (.003)	.167 (.003)	.134 (.004)	.136 (.003)	.172 (.003)	.168 (.005)	.190 (.004)	.158 (.003)	.070 (.003)	.095 (.003)	.169 (.003)	.185 (.004)	.163 (.003)	.172 (.003)	.226 (.005)	.214 (.004)
	Marriage	w/o CausalFT	.149 (.003)	.142 (.003)	.180 (.003)	.161 (.003)	.159 (.004)	.198 (.003)	.176 (.003)	.193 (.005)	.216 (.004)	.147 (.003)	.130 (.003)	.163 (.003)	.163 (.003)	.166 (.004)	.205 (.003)	.177 (.003)	.219 (.005)	.243 (.004)
		w/ CausalFT	.153 (.003)	.147 (.003)	.186 (.003)	.166 (.003)	.158 (.004)	.204 (.003)	.181 (.003)	.199 (.005)	.223 (.004)	.152 (.003)	.135 (.003)	.169 (.003)	.168 (.003)	.171 (.004)	.211 (.003)	.182 (.003)	.226 (.005)	.250 (.004)
	Education	w/o CausalFT	.147 (.003)	.123 (.003)	.087 (.003)	.158 (.003)	.136 (.004)	.097 (.003)	.171 (.003)	.151 (.005)	.119 (.004)	.146 (.003)	.109 (.003)	.083 (.003)	.157 (.003)	.125 (.004)	.096 (.003)	.166 (.003)	.142 (.005)	.113 (.004)
		w/ CausalFT	.152 (.003)	.127 (.003)	.090 (.003)	.163 (.003)	.140 (.004)	.100 (.003)	.176 (.003)	.156 (.005)	.123 (.004)	.151 (.003)	.113 (.003)	.086 (.003)	.162 (.003)	.129 (.004)	.099 (.003)	.171 (.003)	.147 (.005)	.117 (.004)
Crime	*Race*	w/o CausalFT	.160 (.003)	.108 (.003)	.090 (.003)	.175 (.003)	.133 (.004)	.109 (.003)	.186 (.003)	.169 (.005)	.250 (.004)	.158 (.003)	.099 (.003)	.087 (.003)	.166 (.003)	.147 (.004)	.126 (.003)	.185 (.003)	.211 (.005)	.268 (.004)
		w/ CausalFT	.159 (.003)	.112 (.003)	.093 (.003)	.180 (.003)	.137 (.004)	.113 (.003)	.191 (.003)	.174 (.005)	.258 (.004)	.163 (.003)	.102 (.003)	.090 (.003)	.171 (.003)	.152 (.004)	.130 (.003)	.190 (.003)	.218 (.005)	.276 (.004)
	Gender	w/o CausalFT	.163 (.003)	.071 (.003)	.086 (.003)	.181 (.003)	.119 (.004)	.131 (.003)	.190 (.003)	.196 (.005)	.197 (.004)	.161 (.003)	.089 (.003)	.101 (.003)	.175 (.003)	.136 (.004)	.173 (.003)	.190 (.003)	.216 (.005)	.208 (.004)
		w/ CausalFT	.168 (.003)	.073 (.003)	.089 (.003)	.186 (.003)	.123 (.004)	.135 (.003)	.195 (.003)	.202 (.005)	.203 (.004)	.166 (.003)	.092 (.003)	.104 (.003)	.180 (.003)	.140 (.004)	.178 (.003)	.195 (.003)	.223 (.005)	.214 (.004)
German	*Gender*	w/o CausalFT	.162 (.003)	.129 (.003)	.107 (.003)	.176 (.003)	.148 (.004)	.125 (.003)	.183 (.003)	.216 (.005)	.183 (.004)	.160 (.003)	.113 (.003)	.093 (.003)	.169 (.003)	.168 (.004)	.127 (.003)	.181 (.003)	.231 (.005)	.224 (.004)
		w/ CausalFT	.167 (.003)	.133 (.003)	.106 (.003)	.181 (.003)	.153 (.004)	.129 (.003)	.188 (.003)	.223 (.005)	.189 (.004)	.165 (.003)	.117 (.003)	.096 (.003)	.174 (.003)	.173 (.004)	.131 (.003)	.186 (.003)	.238 (.005)	.231 (.004)
	Age	w/o CausalFT	.168 (.003)	.103 (.003)	.134 (.003)	.181 (.003)	.177 (.004)	.164 (.003)	.183 (.003)	.211 (.005)	.183 (.004)	.166 (.003)	.126 (.003)	.127 (.003)	.172 (.003)	.190 (.004)	.166 (.003)	.185 (.003)	.252 (.005)	.208 (.004)
		w/ CausalFT	.173 (.003)	.107 (.003)	.139 (.003)	.186 (.003)	.183 (.004)	.170 (.003)	.188 (.003)	.218 (.005)	.189 (.004)	.171 (.003)	.130 (.003)	.132 (.003)	.177 (.003)	.196 (.004)	.172 (.003)	.190 (.003)	.260 (.005)	.215 (.004)

and compare it against the vanilla generators without CausalFT. This evaluation is based on four different AI models, eight datasets, 2-3 possible sensitive features each, and three metrics, with a total of 324 comparison cases per model. For each comparison, we apply the statistical test and effect size mentioned in Section 4.

5.1.2 Results. From Tables 2, 3, 4, and 5, clearly, we see that CausalFT achieves remarkable improvement over the state-of-the-art generators on all models, finding considerably more fairness bugs in up to 93% of the cases (1209 out of 1296) with statistical significance across all tested models. The above results are consistent regardless of the generator, datasets, and individual/group fairness metrics. In particular, while the figures shown from the metrics might seem small, considering the number of unique test samples generated in our experiments (i.e., 10,000), the practical difference is large. For example, when *Gender* is the concerned sensitive feature under the Dutch dataset, using NeuronFair with CausalFT leads to a mean 0.344 IDI ratio against the mean of 0.302 without CausalFT on DNN_6. If we consider the 10,000 unique test samples generated, this means that the former actually finds 3440 individual discriminatory instances while the latter only finds 3020 ones—pairing with CausalFT can, on average, reveal 420 more unique fairness bugs which is a practically significant improvement. Surprisingly, not only the individual fairness but also the group fairness can be considerably improved by CausalFT, which further suggests the non-trivial correlation between individual and group fairness metrics. All the above demonstrate the benefit of extracting the most causally relevant non-sensitive features to the concerned sensitive feature and injecting such a relationship to guide the perturbation during fairness testing.

In addition, as shown in Figure 4, we see that among the six evaluated base generators (ADF, EIDIG, NeuronFair, SG, ExpGA, and Bream), integrating CausalFT with NeuronFair generally achieves the best performance in revealing fairness bugs across the cases. This is because NeuronFair uses neuron-level adversarial perturbation strategies that directly manipulate hidden model activations to generate fairness-critical inputs. When combined with CausalFT's causality-guided test cases generation, the ability of NeuronFair to explore hidden activations can be amplified. This suggests



(a) DNN_5 (b) DNN_6 (c) Logistic Regression (d) Random Forest

Fig. 4. Comparison of the Scott-Knott test ranks when pairing CausalFT with different generators on all models and cases.

that developers aiming for the best fairness bug discovery ability should prioritize integrating CausalFT with NeuronFair. Overall, we say:

> ***RQ*$_1$*:*** CausalFT *can considerably improve state-of-the-art generators in finding fairness bugs on approximately* 93% *of the cases. Working with* NeuronFair *have enabled* CausalFT *to achieve the generally best results*

5.2 Benefits of Causality Analysis in CausalFT

5.2.1 Method. Since to the best of our knowledge, we cannot find other work that also considers non-sensitive features to manipulate the perturbation for fairness testing, in **RQ**$_2$, we compare CausalFT with FairRF [81], which is a recent work that involves certain steps of similar goal as CausalFT. Briefly, FairRF aims to find some most correlated non-sensitive features to the concerned sensitive one for training the DNN in order to mitigate fairness bugs. Although it targets fairness mitigation rather than fairness testing, FairRF is similar to CausalFT in the sense that it can also rank the non-sensitive feature with respect to the concerned sensitive feature. However, it performs such ranking via correlation analysis rather than understanding the causality as CausalFT does. As such, to compare in our fairness testing context, we only leverage the way that FairRF ranks the non-sensitive features and use the most relevant one to be paired together with the concerned sensitive feature. The method that injects such information into the perturbation is the same as the CausalFT. The same settings for **RQ**$_1$ are used, leading to 324 cases per model.

5.2.2 Results. The results are shown in Tables 6, 7, 8 and 9. In all models, we can see that under six concerned sensitive features, CausalFT and FairRF select the same non-sensitive feature as the most relevant one to its sensitive counterpart, hence they produce the same testing results. This, together with the statistically insignificant comparisons, reveals that both perform similarly in approximately 34% cases (432 out of 1296) across all tested models. However, on most of the other datasets/concerned sensitive features, they make different choices. In general, CausalFT again achieves significantly better outcomes by revealing statistically more fairness bugs over 64% cases (831 out of 1296) across all tested models. This serves as evidence that, when analyzing the relationships between the concerned sensitive feature and the other non-sensitive ones for guiding the perturbation in fairness testing, it is important to consider causality beyond simple correlation. These results indicate that causally guided perturbation not only benefits complex neural networks but also provides advantages for classical models such as LR and RF. Thus, we conclude that:

> ***RQ*$_2$*:*** *The causality analysis in* CausalFT *helps revealing more fairness bugs than current correlation analysis on* 64% *of the cases.*

Table 6. For testing DNN_5, comparing CausalFT with FairRF, which ranks non-sensitive attributes with respect to the sensitive one based on correlation analysis, over all cases using the mean (standard deviation) of all fairness metrics under 10 runs. f_c denotes the direct and most causally relevant non-sensitive feature to the concerned sensitive one. The other format is the same as Table 2.

Dataset	f_s	Approach	f_c	ADF			EIDIG			NeuronFair			SG			ExpGA			BREAM		
				IDI ratio	EOD	SPD	IDI ratio	EOD	SPD	IDI ratio	EOD	SPD	IDI ratio	EOD	SPD	IDI ratio	EOD	SPD	IDI ratio	EOD	SPD
Adult	*Gender*	CausalFT	*Relationship*	.284(.003)	.137(.003)	.099(.003)	.301(.003)	.153(.003)	.112(.003)	.361(.004)	.145(.002)	.130(.003)	.248(.003)	.120(.003)	.123(.003)	.249(.003)	.091(.003)	.099(.002)	.266(.003)	.102(.003)	.109(.003)
		FairRF	*Workclass*	.268(.003)	.121(.003)	.091(.003)	.287(.003)	.137(.003)	.105(.003)	.342(.004)	.136(.003)	.122(.003)	.232(.003)	.112(.003)	.116(.003)	.236(.003)	.084(.003)	.093(.003)	.251(.003)	.096(.003)	.102(.003)
	Race	CausalFT	*Native-country*	.355(.003)	.105(.003)	.125(.003)	.379(.003)	.122(.003)	.129(.003)	.401(.003)	.143(.002)	.145(.003)	.250(.003)	.103(.003)	.107(.003)	.253(.003)	.077(.003)	.080(.002)	.309(.003)	.080(.002)	.085(.003)
		FairRF	*Education*	.338(.003)	.098(.003)	.110(.003)	.360(.003)	.112(.003)	.117(.003)	.384(.003)	.133(.003)	.132(.003)	.235(.003)	.095(.003)	.099(.003)	.239(.003)	.071(.003)	.073(.002)	.294(.003)	.073(.003)	.078(.003)
	Age	CausalFT	*Marital-status*	.333(.003)	.103(.002)	.109(.003)	.327(.002)	.104(.003)	.092(.003)	.389(.003)	.192(.004)	.112(.003)	.232(.003)	.121(.003)	.092(.003)	.255(.003)	.053(.003)	.069(.003)	.278(.003)	.062(.002)	.071(.003)
		FairRF	*Marital-status*	.333(.003)	.103(.002)	.109(.003)	.327(.002)	.104(.003)	.092(.003)	.389(.003)	.192(.004)	.112(.003)	.232(.003)	.121(.003)	.092(.003)	.255(.003)	.053(.003)	.069(.003)	.278(.003)	.062(.002)	.071(.003)
Compas	*Gender*	CausalFT	*Previous Convictions*	.363(.003)	.141(.003)	.122(.002)	.362(.003)	.146(.003)	.114(.003)	.419(.003)	.152(.003)	.140(.003)	.231(.003)	.098(.002)	.093(.003)	.252(.003)	.088(.003)	.097(.003)	.312(.003)	.098(.003)	.118(.003)
		FairRF	*Decile-text*	.347(.003)	.126(.003)	.110(.002)	.349(.003)	.131(.003)	.102(.003)	.398(.003)	.142(.003)	.129(.003)	.214(.003)	.088(.002)	.086(.003)	.240(.003)	.082(.003)	.091(.003)	.325(.003)	.103(.003)	.123(.003)
	Race	CausalFT	*Compas-score*	.394(.003)	.131(.002)	.105(.002)	.437(.003)	.150(.003)	.121(.004)	.446(.003)	.165(.002)	.135(.003)	.249(.003)	.087(.002)	.105(.003)	.270(.003)	.067(.003)	.085(.003)	.348(.003)	.089(.003)	.087(.003)
		FairRF	*Decile-text*	.378(.003)	.116(.002)	.094(.002)	.417(.003)	.136(.003)	.111(.004)	.430(.003)	.151(.002)	.122(.003)	.233(.003)	.081(.002)	.099(.003)	.256(.003)	.062(.003)	.079(.003)	.332(.003)	.081(.003)	.080(.003)
Law School	*Gender*	CausalFT	*Lsat*	.329(.003)	.118(.003)	.100(.003)	.334(.003)	.141(.003)	.115(.002)	.379(.002)	.158(.003)	.123(.003)	.224(.003)	.106(.003)	.108(.003)	.236(.003)	.066(.003)	.087(.002)	.291(.003)	.077(.003)	.100(.003)
		FairRF	*Decile1b*	.312(.003)	.106(.003)	.090(.003)	.318(.003)	.127(.003)	.104(.002)	.360(.003)	.146(.003)	.114(.003)	.205(.003)	.095(.003)	.099(.003)	.219(.003)	.060(.003)	.078(.002)	.275(.003)	.069(.003)	.090(.003)
	Race	CausalFT	*Lsat*	.301(.003)	.114(.003)	.089(.003)	.330(.003)	.126(.003)	.105(.003)	.396(.003)	.139(.003)	.126(.003)	.234(.003)	.099(.003)	.098(.003)	.258(.003)	.073(.003)	.069(.003)	.285(.003)	.082(.003)	.089(.002)
		FairRF	*Delice3*	.286(.003)	.111(.003)	.083(.003)	.314(.003)	.119(.003)	.098(.003)	.378(.003)	.132(.003)	.118(.003)	.220(.003)	.093(.003)	.092(.003)	.244(.003)	.067(.003)	.063(.003)	.272(.003)	.074(.003)	.084(.002)
Kdd	*Gender*	CausalFT	*Occupation*	.290(.003)	.129(.003)	.114(.002)	.284(.002)	.137(.003)	.111(.003)	.304(.003)	.160(.003)	.119(.003)	.222(.003)	.090(.003)	.095(.003)	.221(.002)	.077(.002)	.095(.004)	.261(.003)	.081(.003)	.107(.003)
		FairRF	*Housing*	.274(.003)	.115(.003)	.106(.002)	.270(.003)	.123(.003)	.103(.003)	.287(.003)	.146(.003)	.110(.003)	.206(.003)	.081(.003)	.087(.003)	.205(.002)	.070(.002)	.086(.004)	.245(.003)	.072(.003)	.097(.003)
	Race	CausalFT	*Tax-filer-stat*	.293(.003)	.104(.003)	.111(.002)	.291(.003)	.146(.003)	.110(.003)	.309(.003)	.145(.003)	.125(.003)	.242(.003)	.086(.003)	.097(.003)	.221(.003)	.098(.003)	.103(.003)	.244(.003)	.089(.003)	.114(.003)
		FairRF	*Marital-status*	.277(.003)	.096(.003)	.101(.002)	.276(.003)	.132(.003)	.101(.003)	.292(.003)	.131(.003)	.114(.003)	.226(.003)	.079(.003)	.088(.003)	.206(.003)	.090(.003)	.095(.003)	.229(.003)	.081(.003)	.102(.003)
Dutch	*Gender*	CausalFT	*Cur_eco_activity*	.295(.004)	.127(.003)	.097(.003)	.291(.002)	.139(.003)	.105(.003)	.323(.003)	.148(.003)	.119(.003)	.225(.003)	.110(.002)	.107(.003)	.226(.003)	.093(.003)	.087(.003)	.256(.003)	.081(.003)	.096(.003)
		FairRF	*Cur_eco_activity*	.295(.004)	.127(.003)	.097(.003)	.291(.002)	.139(.003)	.105(.003)	.323(.003)	.148(.003)	.119(.003)	.225(.003)	.110(.002)	.107(.003)	.226(.003)	.093(.003)	.087(.003)	.256(.003)	.081(.003)	.096(.003)
	Age	CausalFT	*Economic_status*	.288(.003)	.109(.003)	.119(.003)	.305(.003)	.138(.002)	.113(.003)	.308(.003)	.158(.003)	.135(.003)	.210(.003)	.103(.002)	.107(.003)	.219(.003)	.090(.003)	.096(.003)	.227(.003)	.134(.003)	.113(.003)
		FairRF	*Edu_level*	.273(.003)	.098(.003)	.108(.003)	.288(.003)	.124(.002)	.101(.003)	.292(.003)	.142(.003)	.122(.003)	.198(.003)	.094(.002)	.099(.003)	.206(.003)	.082(.003)	.087(.003)	.212(.003)	.119(.003)	.103(.003)
Credit	*Gender*	CausalFT	*Pay_amt3*	.302(.003)	.107(.003)	.109(.003)	.275(.004)	.116(.003)	.094(.003)	.302(.003)	.146(.003)	.101(.003)	.212(.003)	.080(.003)	.095(.003)	.198(.003)	.056(.003)	.080(.003)	.217(.002)	.066(.003)	.089(.003)
		FairRF	*Pay_amt3*	.302(.003)	.107(.003)	.109(.003)	.275(.004)	.116(.003)	.094(.003)	.302(.003)	.146(.003)	.101(.003)	.212(.003)	.080(.003)	.095(.003)	.198(.003)	.056(.003)	.080(.003)	.217(.002)	.066(.003)	.089(.003)
	Marriage	CausalFT	*Pay_amt1*	.292(.003)	.113(.003)	.082(.004)	.324(.003)	.122(.002)	.090(.003)	.312(.003)	.138(.003)	.097(.003)	.247(.003)	.099(.003)	.086(.003)	.245(.003)	.073(.003)	.068(.003)	.260(.003)	.077(.003)	.072(.002)
		FairRF	*Pay_amt1*	.292(.003)	.113(.003)	.082(.004)	.324(.003)	.122(.002)	.090(.003)	.312(.003)	.138(.003)	.097(.003)	.247(.003)	.099(.003)	.086(.003)	.245(.003)	.073(.003)	.068(.003)	.260(.003)	.077(.003)	.072(.002)
	Education	CausalFT	*Pay_amt1*	.310(.003)	.128(.003)	.145(.003)	.313(.003)	.128(.003)	.138(.003)	.324(.003)	.143(.003)	.120(.003)	.246(.003)	.105(.003)	.093(.003)	.227(.003)	.100(.003)	.120(.003)	.245(.003)	.105(.003)	.112(.003)
		FairRF	*BILL_AMT3*	.296(.003)	.114(.003)	.131(.003)	.298(.003)	.116(.003)	.126(.003)	.310(.003)	.131(.003)	.111(.003)	.232(.003)	.093(.003)	.084(.003)	.214(.003)	.089(.003)	.108(.003)	.232(.003)	.091(.003)	.099(.003)
Crime	*Race*	CausalFT	*Score_text*	.278(.003)	.118(.003)	.090(.003)	.284(.002)	.139(.002)	.094(.002)	.309(.003)	.155(.003)	.108(.002)	.254(.003)	.080(.003)	.102(.003)	.237(.003)	.067(.003)	.083(.003)	.257(.003)	.078(.004)	.099(.002)
		FairRF	*Score_text*	.278(.003)	.118(.003)	.090(.003)	.284(.002)	.139(.002)	.094(.002)	.309(.003)	.155(.003)	.108(.002)	.254(.003)	.080(.003)	.102(.003)	.237(.003)	.067(.003)	.083(.003)	.257(.003)	.078(.004)	.099(.002)
	Gender	CausalFT	*Priors_count*	.315(.003)	.120(.003)	.108(.004)	.318(.002)	.137(.002)	.116(.002)	.332(.002)	.171(.003)	.134(.003)	.249(.003)	.103(.003)	.102(.003)	.256(.002)	.109(.002)	.106(.003)	.273(.003)	.077(.003)	.109(.003)
		FairRF	*V_score_text*	.298(.003)	.108(.003)	.097(.004)	.303(.002)	.125(.002)	.105(.002)	.316(.002)	.158(.003)	.123(.003)	.234(.003)	.093(.003)	.093(.003)	.241(.002)	.098(.002)	.096(.003)	.259(.003)	.069(.003)	.101(.003)
German	*Gender*	CausalFT	*Credit-amount*	.321(.004)	.118(.002)	.105(.003)	.303(.003)	.141(.003)	.145(.003)	.331(.003)	.154(.003)	.149(.002)	.229(.004)	.088(.003)	.074(.002)	.225(.003)	.090(.003)	.078(.002)	.266(.002)	.114(.004)	.092(.003)
		FairRF	*Credit-amount*	.321(.004)	.118(.002)	.105(.003)	.303(.003)	.141(.003)	.145(.003)	.331(.003)	.154(.003)	.149(.002)	.229(.004)	.088(.003)	.074(.002)	.225(.003)	.090(.003)	.078(.002)	.266(.002)	.114(.004)	.092(.003)
	Age	CausalFT	*Existing-credits*	.260(.003)	.131(.004)	.101(.003)	.283(.003)	.146(.003)	.122(.003)	.310(.003)	.143(.003)	.127(.003)	.264(.003)	.116(.003)	.116(.003)	.240(.003)	.109(.003)	.106(.003)	.257(.003)	.091(.003)	.103(.003)
		FairRF	*Duration*	.247(.003)	.118(.004)	.093(.003)	.271(.003)	.151(.003)	.114(.003)	.295(.003)	.129(.003)	.118(.003)	.248(.003)	.110(.003)	.108(.003)	.228(.003)	.104(.003)	.099(.003)	.243(.003)	.086(.003)	.094(.003)

Table 7. When testing DNN_6, comparing CausalFT with FairRF. The other format is the same as Table 6.

Dataset	f_s	Approach	f_c	ADF			EIDIG			NeuronFair			SG			ExpGA			BREAM		
				IDI ratio	EOD	SPD	IDI ratio	EOD	SPD	IDI ratio	EOD	SPD	IDI ratio	EOD	SPD	IDI ratio	EOD	SPD	IDI ratio	EOD	SPD
Adult	*Gender*	CausalFT	*Relationship*	.301(.003)	.142(.003)	.103(.003)	.316(.003)	.169(.003)	.120(.003)	.365(.004)	.154(.002)	.137(.003)	.262(.003)	.127(.003)	.116(.003)	.274(.003)	.092(.003)	.104(.002)	.284(.003)	.109(.003)	.121(.003)
		FairRF	*Workclass*	.293(.003)	.109(.003)	.085(.003)	.283(.003)	.126(.003)	.078(.004)	.339(.002)	.143(.002)	.107(.001)	.222(.003)	.067(.003)	.085(.002)	.257(.004)	.065(.003)	.102(.003)	.263(.002)	.075(.003)	.109(.003)
	Race	CausalFT	*Native-country*	.382(.003)	.113(.003)	.132(.003)	.396(.003)	.131(.003)	.145(.003)	.429(.003)	.154(.002)	.155(.003)	.249(.003)	.102(.003)	.116(.003)	.251(.003)	.079(.003)	.079(.001)	.316(.003)	.086(.002)	.097(.003)
		FairRF	*Education*	.377(.003)	.105(.002)	.068(.002)	.341(.002)	.109(.003)	.134(.002)	.341(.003)	.148(.002)	.139(.003)	.165(.003)	.057(.003)	.073(.003)	.217(.003)	.075(.003)	.072(.003)	.307(.003)	.074(.003)	.092(.003)
	Age	CausalFT	*Marital-status*	.353(.003)	.111(.002)	.117(.003)	.357(.002)	.113(.004)	.104(.003)	.388(.003)	.209(.004)	.122(.003)	.251(.003)	.122(.003)	.098(.003)	.274(.003)	.056(.003)	.069(.003)	.308(.003)	.063(.002)	.074(.003)
		FairRF	*Marital-status*	.353(.003)	.111(.002)	.117(.003)	.357(.002)	.113(.004)	.104(.003)	.388(.003)	.209(.004)	.122(.003)	.251(.002)	.122(.003)	.098(.003)	.274(.003)	.056(.003)	.069(.003)	.308(.003)	.063(.002)	.074(.003)
Compas	*Gender*	CausalFT	*Previous Convictions*	.386(.003)	.144(.003)	.130(.002)	.394(.003)	.157(.003)	.124(.003)	.419(.003)	.163(.003)	.153(.003)	.247(.003)	.103(.002)	.094(.003)	.272(.003)	.092(.003)	.105(.003)	.335(.003)	.110(.003)	.126(.003)
		FairRF	*Decile-text*	.362(.003)	.115(.003)	.077(.003)	.377(.003)	.096(.003)	.101(.003)	.383(.002)	.157(.003)	.111(.002)	.211(.003)	.056(.003)	.068(.003)	.249(.003)	.064(.003)	.173(.003)	.291(.003)	.106(.003)	.132(.002)
	Race	CausalFT	*Compas-score*	.426(.003)	.138(.002)	.113(.002)	.435(.003)	.158(.003)	.130(.004)	.482(.003)	.175(.002)	.140(.003)	.272(.003)	.094(.002)	.112(.003)	.297(.003)	.070(.003)	.087(.003)	.363(.003)	.083(.003)	.095(.003)
		FairRF	*Decile-text*	.405(.003)	.112(.003)	.104(.003)	.411(.002)	.137(.003)	.121(.003)	.432(.003)	.142(.003)	.125(.003)	.227(.002)	.064(.003)	.081(.002)	.263(.003)	.066(.002)	.069(.003)	.323(.003)	.079(.004)	.083(.003)
Law School	*Gender*	CausalFT	*Lsat*	.334(.003)	.128(.003)	.107(.003)	.356(.003)	.150(.003)	.125(.002)	.398(.002)	.166(.003)	.133(.003)	.239(.003)	.107(.003)	.121(.003)	.254(.003)	.067(.003)	.095(.002)	.313(.003)	.082(.003)	.107(.003)
		FairRF	*Decile1b*	.301(.003)	.094(.003)	.086(.003)	.322(.004)	.128(.003)	.104(.002)	.352(.003)	.142(.004)	.127(.003)	.198(.002)	.062(.003)	.117(.005)	.216(.003)	.072(.003)	.083(.004)	.247(.002)	.074(.004)	.093(.003)
	Race	CausalFT	*Lsat*	.322(.003)	.114(.003)	.092(.003)	.333(.003)	.138(.003)	.113(.003)	.385(.003)	.149(.003)	.128(.003)	.251(.003)	.102(.002)	.111(.003)	.273(.003)	.079(.003)	.079(.003)	.308(.003)	.079(.003)	.096(.002)
		FairRF	*Delice3*	.319(.003)	.105(.003)	.090(.003)	.325(.003)	.136(.003)	.107(.002)	.367(.003)	.141(.003)	.113(.002)	.232(.003)	.081(.002)	.077(.003)	.267(.003)	.108(.002)	.071(.003)	.302(.003)	.116(.003)	.085(.003)
Kdd	*Gender*	CausalFT	*Occupation*	.308(.003)	.134(.003)	.116(.002)	.309(.002)	.148(.003)	.118(.003)	.324(.003)	.170(.003)	.124(.003)	.251(.003)	.097(.002)	.103(.003)	.246(.002)	.079(.002)	.102(.004)	.260(.003)	.086(.003)	.119(.003)
		FairRF	*Housing*	.276(.003)	.114(.003)	.114(.003)	.289(.003)	.124(.002)	.097(.003)	.318(.002)	.152(.003)	.111(.003)	.218(.003)	.071(.003)	.087(.003)	.227(.003)	.076(.003)	.093(.003)	.253(.002)	.077(.003)	.108(.003)
	Race	CausalFT	*Tax-filer-stat*	.311(.003)	.116(.003)	.121(.002)	.317(.003)	.142(.003)	.119(.003)	.332(.003)	.152(.003)	.143(.003)	.242(.003)	.097(.003)	.103(.003)	.245(.003)	.104(.003)	.111(.003)	.262(.003)	.092(.003)	.120(.003)
		FairRF	*Marital-status*	.291(.002)	.112(.003)	.087(.003)	.304(.003)	.120(.002)	.089(.003)	.314(.003)	.132(.002)	.114(.003)	.221(.002)	.065(.002)	.071(.003)	.237(.004)	.071(.003)	.077(.004)	.243(.003)	.069(.003)	.066(.002)
Dutch	*Gender*	CausalFT	*Cur_eco_activity*	.319(.004)	.134(.003)	.100(.003)	.322(.002)	.149(.003)	.112(.003)	.344(.003)	.162(.003)	.127(.003)	.233(.003)	.089(.002)	.091(.002)	.238(.003)	.095(.003)	.095(.003)	.252(.003)	.087(.003)	.102(.003)
		FairRF	*Cur_eco_activity*	.319(.004)	.134(.003)	.100(.003)	.322(.002)	.149(.003)	.112(.003)	.344(.003)	.162(.003)	.127(.003)	.233(.003)	.089(.002)	.091(.002)	.238(.003)	.095(.003)	.095(.003)	.252(.003)	.087(.003)	.102(.003)
	Age	CausalFT	*Economic_status*	.313(.003)	.123(.003)	.131(.003)	.321(.003)	.146(.002)	.124(.003)	.334(.003)	.164(.003)	.144(.003)	.235(.003)	.105(.003)	.112(.003)	.233(.003)	.094(.003)	.103(.003)	.247(.003)	.142(.003)	.122(.003)
		FairRF	*Edu_level*	.297(.003)	.102(.003)	.110(.003)	.301(.003)	.129(.003)	.113(.003)	.314(.003)	.131(.003)	.117(.003)	.212(.002)	.068(.002)	.088(.003)	.223(.003)	.068(.003)	.092(.003)	.235(.002)	.139(.001)	.101(.003)
Credit	*Gender*	CausalFT	*Pay_amt3*	.292(.003)	.114(.003)	.108(.003)	.299(.004)	.124(.003)	.104(.003)	.316(.003)	.157(.003)	.111(.003)	.219(.003)	.086(.003)	.101(.003)	.215(.003)	.054(.003)	.088(.003)	.237(.002)	.069(.003)	.093(.003)
		FairRF	*Pay_amt3*	.292(.003)	.114(.003)	.108(.003)	.299(.004)	.124(.003)	.104(.003)	.316(.003)	.157(.003)	.111(.003)	.219(.003)	.086(.003)	.101(.003)	.215(.003)	.054(.003)	.088(.003)	.237(.002)	.069(.003)	.093(.003)
	Marriage	CausalFT	*Pay_amt1*	.308(.003)	.118(.003)	.089(.004)	.317(.003)	.135(.002)	.093(.003)	.324(.003)	.149(.003)	.102(.003)	.263(.003)	.101(.003)	.092(.003)	.261(.003)	.073(.003)	.067(.003)	.284(.003)	.079(.003)	.077(.002)
		FairRF	*Pay_amt1*	.308(.003)	.118(.003)	.089(.004)	.317(.003)	.135(.002)	.093(.003)	.324(.003)	.149(.003)	.102(.003)	.263(.003)	.101(.003)	.092(.003)	.261(.003)	.073(.003)	.067(.003)	.284(.003)	.079(.003)	.077(.002)
	Education	CausalFT	*Pay_amt1*	.324(.003)	.136(.003)	.154(.003)	.327(.003)	.135(.003)	.142(.003)	.343(.003)	.156(.003)	.119(.003)	.242(.003)	.102(.003)	.097(.003)	.245(.003)	.110(.003)	.127(.003)	.263(.003)	.113(.003)	.121(.003)
		FairRF	*BILL_AMT3*	.310(.003)	.109(.004)	.104(.004)	.311(.003)	.128(.003)	.109(.003)	.331(.003)	.142(.003)	.114(.003)	.228(.003)	.063(.003)	.069(.002)	.232(.003)	.068(.003)	.072(.003)	.248(.003)	.082(.002)	.084(.003)
Crime	*Race*	CausalFT	*Score_text*	.307(.003)	.126(.003)	.100(.003)	.309(.002)	.150(.002)	.103(.002)	.331(.003)	.169(.003)	.115(.002)	.246(.003)	.082(.003)	.102(.003)	.248(.003)	.071(.003)	.087(.003)	.269(.003)	.082(.004)	.097(.002)
		FairRF	*Score_text*	.307(.003)	.126(.003)	.100(.003)	.309(.002)	.150(.002)	.103(.002)	.331(.003)	.169(.003)	.115(.002)	.246(.003)	.082(.003)	.102(.003)	.248(.003)	.071(.003)	.087(.003)	.269(.003)	.082(.004)	.097(.002)
	Gender	CausalFT	*Priors_count*	.332(.003)	.131(.003)	.114(.004)	.330(.002)	.145(.002)	.126(.002)	.352(.002)	.168(.003)	.136(.003)	.271(.003)	.114(.003)	.102(.003)	.278(.002)	.117(.002)	.113(.003)	.294(.003)	.081(.003)	.119(.003)
		FairRF	*V_score_text*	.311(.003)	.117(.004)	.104(.003)	.312(.003)	.121(.003)	.105(.002)	.336(.003)	.134(.003)	.112(.002)	.244(.004)	.103(.003)	.091(.003)	.252(.004)	.104(.002)	.097(.003)	.278(.003)	.083(.003)	.107(.003)
German	*Gender*	CausalFT	*Credit-amount*	.317(.004)	.129(.002)	.112(.003)	.324(.003)	.139(.003)	.145(.003)	.333(.003)	.150(.003)	.147(.002)	.241(.004)	.097(.003)	.082(.002)	.246(.003)	.094(.003)	.084(.002)	.264(.002)	.121(.004)	.095(.003)
		FairRF	*Credit-amount*	.317(.004)	.129(.002)	.112(.003)	.324(.003)	.139(.003)	.145(.003)	.333(.003)	.150(.003)	.147(.002)	.241(.004)	.097(.003)	.082(.002)	.246(.003)	.094(.003)	.084(.002)	.264(.002)	.121(.004)	.095(.003)
	Age	CausalFT	*Existing-credits*	.284(.003)	.142(.004)	.113(.003)	.311(.003)	.154(.003)	.127(.003)	.329(.003)	.156(.003)	.134(.003)	.258(.003)	.124(.003)	.129(.003)	.259(.003)	.125(.003)	.111(.003)	.280(.003)	.103(.003)	.114(.003)
		FairRF	*Duration*	.269(.003)	.126(.003)	.084(.003)	.305(.003)	.128(.002)	.094(.002)	.317(.003)	.143(.002)	.108(.003)	.247(.003)	.083(.003)	.078(.002)	.246(.003)	.094(.003)	.087(.002)	.269(.003)	.081(.003)	.083(.003)

5.3 Effect of Relaxation and Invalidity Repair

5.3.1 Method. For **RQ**$_3$, we count the number of pairs generated with the following phases/variants of CausalFT for generating 10, 000 unique test samples[10]:

- **w/o R:** The number of unique pairs generated for the case without relaxation, i.e., no CausalFT but directly using the base generator.
- **w/ R:** The number of unique pairs generated with CausalFT before invalidity repair.

[10] Note that we only consider the pairs where at least one of the samples matches with the original testing dataset.

Table 8. When testing LR, comparing CausalFT with FairRF. The other format is the same as Table 6.

Dataset	f_s	Approach	f_c	ADF			EIDIG			NeuronFair			SG			ExpGA			BREAM		
				IDI ratio	EOD	SPD	IDI ratio	EOD	SPD	IDI ratio	EOD	SPD	IDI ratio	EOD	SPD	IDI ratio	EOD	SPD	IDI ratio	EOD	SPD
Adult	Gender	CausalFT	Relationship	.180 (.003)	.080 (.003)	.206 (.003)	.189 (.003)	.112 (.002)	.193 (.001)	.191 (.002)	.096 (.002)	.152 (.002)	.156 (.004)	.127 (.004)	.180 (.006)	.188 (.003)	.150 (.001)	.105 (.004)	.175 (.003)	.185 (.003)	.088 (.003)
		FairRF	Workclass	.165 (.003)	.079 (.003)	.191 (.003)	.174 (.003)	.098 (.002)	.178 (.001)	.176 (.002)	.082 (.002)	.138 (.002)	.142 (.004)	.113 (.004)	.166 (.006)	.173 (.003)	.136 (.001)	.100 (.004)	.160 (.003)	.170 (.003)	.074 (.003)
	Race	CausalFT	Native-country	.178 (.003)	.114 (.005)	.217 (.004)	.169 (.003)	.123 (.004)	.073 (.003)	.176 (.002)	.210 (.003)	.173 (.002)	.174 (.002)	.127 (.004)	.180 (.006)	.182 (.002)	.180 (.003)	.128 (.003)	.170 (.003)	.146 (.002)	.079 (.003)
		FairRF	Education	.163 (.003)	.100 (.005)	.202 (.004)	.154 (.003)	.109 (.004)	.068 (.003)	.161 (.002)	.195 (.003)	.158 (.002)	.159 (.002)	.113 (.004)	.165 (.006)	.167 (.002)	.165 (.003)	.134 (.003)	.155 (.003)	.131 (.002)	.084 (.003)
	Age	CausalFT	Marital-status	.169 (.002)	.159 (.003)	.242 (.003)	.178 (.002)	.146 (.003)	.236 (.003)	.196 (.002)	.149 (.003)	.246 (.003)	.143 (.002)	.164 (.003)	.227 (.003)	.159 (.002)	.145 (.003)	.240 (.003)	.164 (.002)	.149 (.003)	.245 (.003)
		FairRF	Marital-status	.169 (.002)	.159 (.003)	.242 (.003)	.178 (.002)	.146 (.003)	.236 (.003)	.196 (.002)	.149 (.003)	.246 (.003)	.143 (.002)	.164 (.003)	.227 (.003)	.159 (.002)	.145 (.003)	.240 (.003)	.164 (.002)	.149 (.003)	.245 (.003)
Compas	Gender	CausalFT	Previous Convictions	.179 (.002)	.206 (.004)	.117 (.003)	.169 (.002)	.074 (.002)	.243 (.002)	.191 (.002)	.197 (.002)	.281 (.002)	.158 (.003)	.127 (.004)	.180 (.004)	.180 (.003)	.148 (.004)	.121 (.004)	.179 (.003)	.195 (.001)	.191 (.002)
		FairRF	Decile-text	.165 (.002)	.191 (.004)	.103 (.003)	.155 (.002)	.069 (.002)	.228 (.002)	.177 (.002)	.202 (.002)	.266 (.002)	.144 (.003)	.113 (.004)	.166 (.004)	.166 (.003)	.134 (.004)	.107 (.004)	.165 (.003)	.180 (.001)	.177 (.002)
	Race	CausalFT	Compas-score	.173 (.003)	.107 (.002)	.080 (.004)	.180 (.003)	.161 (.004)	.109 (.002)	.195 (.003)	.148 (.004)	.127 (.004)	.152 (.003)	.127 (.004)	.180 (.003)	.160 (.003)	.089 (.003)	.215 (.002)	.170 (.002)	.098 (.002)	.162 (.001)
		FairRF	Decile-text	.159 (.003)	.093 (.002)	.080 (.004)	.166 (.003)	.147 (.004)	.095 (.002)	.181 (.003)	.134 (.004)	.113 (.004)	.138 (.003)	.113 (.004)	.166 (.003)	.146 (.003)	.084 (.003)	.201 (.002)	.156 (.002)	.084 (.002)	.148 (.001)
Law School	Gender	CausalFT	Lsat	.162 (.003)	.074 (.004)	.117 (.004)	.177 (.003)	.110 (.004)	.085 (.004)	.178 (.003)	.157 (.004)	.195 (.002)	.161 (.002)	.148 (.002)	.238 (.003)	.155 (.003)	.179 (.002)	.239 (.001)	.165 (.003)	.092 (.005)	.174 (.004)
		FairRF	Decile1b	.148 (.003)	.073 (.004)	.103 (.004)	.163 (.003)	.096 (.004)	.071 (.004)	.164 (.003)	.143 (.004)	.181 (.002)	.147 (.002)	.134 (.002)	.224 (.003)	.141 (.003)	.165 (.002)	.225 (.001)	.151 (.003)	.078 (.005)	.160 (.004)
	Race	CausalFT	Lsat	.175 (.002)	.133 (.004)	.119 (.004)	.178 (.002)	.167 (.004)	.233 (.004)	.181 (.003)	.139 (.004)	.183 (.003)	.159 (.003)	.201 (.002)	.191 (.002)	.179 (.003)	.215 (.001)	.211 (.002)	.169 (.002)	.204 (.004)	.239 (.005)
		FairRF	Delice3	.161 (.002)	.119 (.004)	.105 (.004)	.164 (.002)	.153 (.004)	.219 (.004)	.167 (.003)	.125 (.004)	.169 (.003)	.145 (.003)	.187 (.002)	.177 (.002)	.165 (.003)	.201 (.001)	.197 (.002)	.155 (.002)	.190 (.004)	.225 (.005)
Kdd	Gender	CausalFT	Occupation	.168 (.003)	.130 (.004)	.069 (.001)	.175 (.003)	.177 (.004)	.078 (.001)	.180 (.003)	.076 (.002)	.225 (.004)	.170 (.003)	.116 (.002)	.187 (.003)	.169 (.003)	.099 (.004)	.225 (.004)	.171 (.004)	.096 (.004)	.154 (.001)
		FairRF	Housing	.153 (.003)	.115 (.004)	.072 (.001)	.160 (.003)	.162 (.004)	.073 (.001)	.165 (.003)	.068 (.002)	.210 (.004)	.155 (.003)	.102 (.002)	.172 (.003)	.154 (.003)	.086 (.004)	.210 (.004)	.156 (.004)	.083 (.004)	.140 (.001)
	Race	CausalFT	Tax-filer-stat	.178 (.002)	.207 (.002)	.275 (.003)	.171 (.003)	.110 (.003)	.160 (.004)	.177 (.002)	.093 (.002)	.236 (.003)	.166 (.002)	.101 (.002)	.205 (.003)	.171 (.003)	.097 (.002)	.229 (.004)	.169 (.002)	.100 (.002)	.235 (.003)
		FairRF	Marital-status	.163 (.002)	.192 (.002)	.260 (.003)	.156 (.003)	.105 (.003)	.145 (.004)	.162 (.002)	.080 (.002)	.221 (.003)	.151 (.002)	.088 (.002)	.190 (.003)	.156 (.003)	.084 (.002)	.214 (.004)	.154 (.002)	.087 (.002)	.220 (.003)
Dutch	Gender	CausalFT	Cur_eco_activity	.174 (.002)	.109 (.004)	.258 (.003)	.164 (.002)	.076 (.001)	.189 (.003)	.175 (.003)	.141 (.003)	.232 (.003)	.164 (.002)	.169 (.004)	.228 (.003)	.183 (.003)	.102 (.004)	.196 (.002)	.177 (.002)	.074 (.002)	.102 (.004)
		FairRF	Cur_eco_activity	.174 (.002)	.109 (.004)	.258 (.003)	.164 (.002)	.076 (.001)	.189 (.003)	.175 (.003)	.141 (.003)	.232 (.003)	.164 (.002)	.169 (.004)	.228 (.003)	.183 (.003)	.102 (.004)	.196 (.002)	.177 (.002)	.074 (.002)	.102 (.004)
	Age	CausalFT	Economic_status	.164 (.003)	.074 (.001)	.093 (.003)	.169 (.003)	.176 (.002)	.183 (.004)	.194 (.002)	.191 (.004)	.082 (.002)	.175 (.003)	.218 (.003)	.243 (.002)	.185 (.003)	.208 (.003)	.253 (.002)	.187 (.003)	.114 (.005)	.107 (.004)
		FairRF	Edu_level	.149 (.003)	.073 (.001)	.080 (.003)	.154 (.003)	.161 (.002)	.168 (.004)	.179 (.002)	.176 (.004)	.070 (.002)	.160 (.003)	.203 (.003)	.228 (.002)	.170 (.003)	.193 (.003)	.238 (.002)	.172 (.003)	.100 (.005)	.094 (.004)
Credit	Gender	CausalFT	Pay_amt3	.171 (.002)	.074 (.001)	.120 (.003)	.175 (.002)	.093 (.002)	.067 (.004)	.166 (.002)	.077 (.001)	.270 (.004)	.160 (.002)	.151 (.003)	.148 (.002)	.182 (.003)	.162 (.003)	.090 (.003)	.170 (.002)	.131 (.003)	.188 (.002)
		FairRF	Pay_amt3	.171 (.002)	.074 (.001)	.120 (.003)	.175 (.002)	.093 (.002)	.067 (.004)	.166 (.002)	.077 (.001)	.270 (.004)	.160 (.002)	.151 (.003)	.148 (.002)	.182 (.003)	.162 (.003)	.090 (.003)	.170 (.002)	.131 (.003)	.188 (.002)
	Marriage	CausalFT	Pay_amt1	.164 (.001)	.132 (.004)	.162 (.001)	.188 (.003)	.091 (.003)	.120 (.002)	.191 (.003)	.120 (.003)	.170 (.003)	.168 (.002)	.219 (.001)	.264 (.002)	.151 (.003)	.201 (.003)	.241 (.002)	.169 (.001)	.226 (.002)	.259 (.002)
		FairRF	Pay_amt1	.201 (.002)	.168 (.003)	.256 (.003)	.210 (.002)	.162 (.003)	.263 (.003)	.226 (.002)	.163 (.003)	.262 (.003)	.174 (.002)	.181 (.003)	.242 (.003)	.189 (.002)	.189 (.003)	.253 (.003)	.186 (.002)	.158 (.003)	.232 (.003)
	Education	CausalFT	Pay_amt1	.147 (.003)	.086 (.002)	.162 (.002)	.170 (.003)	.071 (.005)	.215 (.004)	.173 (.003)	.219 (.005)	.141 (.003)	.157 (.002)	.143 (.002)	.107 (.003)	.162 (.002)	.177 (.002)	.127 (.003)	.174 (.003)	.159 (.004)	.056 (.002)
		FairRF	BILL_AMT3	.132 (.003)	.078 (.002)	.148 (.002)	.155 (.003)	.075 (.005)	.200 (.004)	.158 (.003)	.204 (.005)	.128 (.003)	.142 (.002)	.130 (.002)	.095 (.003)	.147 (.002)	.162 (.002)	.114 (.003)	.159 (.003)	.145 (.004)	.060 (.002)
Crime	Race	CausalFT	Score_text	.147 (.003)	.086 (.002)	.162 (.002)	.170 (.003)	.071 (.005)	.215 (.004)	.173 (.003)	.219 (.005)	.141 (.003)	.157 (.002)	.143 (.002)	.107 (.003)	.162 (.002)	.177 (.002)	.127 (.003)	.174 (.003)	.159 (.004)	.056 (.002)
		FairRF	Score_text	.147 (.003)	.086 (.002)	.162 (.002)	.170 (.003)	.071 (.005)	.215 (.004)	.173 (.003)	.219 (.005)	.141 (.003)	.157 (.002)	.143 (.002)	.107 (.003)	.162 (.002)	.177 (.002)	.127 (.003)	.174 (.003)	.159 (.004)	.056 (.002)
	Gender	CausalFT	Priors_count	.168 (.002)	.149 (.003)	.125 (.002)	.185 (.003)	.102 (.003)	.234 (.002)	.194 (.002)	.069 (.001)	.206 (.002)	.164 (.002)	.101 (.002)	.150 (.002)	.174 (.002)	.111 (.002)	.158 (.002)	.184 (.003)	.122 (.002)	.290 (.002)
		FairRF	V_score_text	.153 (.002)	.135 (.003)	.113 (.002)	.170 (.003)	.107 (.003)	.239 (.002)	.179 (.002)	.063 (.001)	.192 (.002)	.150 (.002)	.092 (.002)	.138 (.002)	.159 (.002)	.101 (.002)	.146 (.002)	.169 (.003)	.112 (.002)	.275 (.002)
German	Gender	CausalFT	Credit-amount	.178 (.002)	.108 (.003)	.084 (.004)	.176 (.002)	.171 (.003)	.272 (.002)	.180 (.003)	.108 (.003)	.203 (.004)	.171 (.003)	.127 (.004)	.180 (.003)	.170 (.003)	.138 (.003)	.219 (.002)	.169 (.002)	.212 (.005)	.259 (.002)
		FairRF	Credit-amount	.178 (.002)	.108 (.003)	.084 (.004)	.176 (.002)	.171 (.003)	.272 (.002)	.180 (.003)	.108 (.003)	.203 (.004)	.171 (.003)	.127 (.004)	.180 (.003)	.170 (.003)	.138 (.003)	.219 (.002)	.169 (.002)	.212 (.005)	.259 (.002)
	Age	CausalFT	Existing-credits	.178 (.002)	.096 (.002)	.132 (.004)	.174 (.003)	.148 (.005)	.122 (.003)	.190 (.002)	.195 (.004)	.151 (.005)	.161 (.002)	.127 (.004)	.180 (.006)	.175 (.002)	.168 (.002)	.107 (.002)	.167 (.003)	.173 (.003)	.131 (.004)
		FairRF	Duration	.162 (.002)	.088 (.002)	.130 (.004)	.159 (.003)	.135 (.005)	.110 (.003)	.175 (.002)	.180 (.004)	.138 (.005)	.147 (.002)	.115 (.004)	.167 (.006)	.160 (.002)	.173 (.002)	.095 (.002)	.152 (.003)	.158 (.003)	.128 (.004)

Table 9. When testing RF, comparing CausalFT with FairRF. The other format is the same as Table 6.

Dataset	f_s	Approach	f_c	ADF			EIDIG			NeuronFair			SG			ExpGA			BREAM		
				IDI ratio	EOD	SPD	IDI ratio	EOD	SPD	IDI ratio	EOD	SPD	IDI ratio	EOD	SPD	IDI ratio	EOD	SPD	IDI ratio	EOD	SPD
Adult	Gender	CausalFT	Relationship	.165 (.003)	.079 (.004)	.129 (.003)	.171 (.003)	.145 (.004)	.167 (.003)	.186 (.003)	.159 (.005)	.189 (.004)	.159 (.003)	.106 (.003)	.080 (.003)	.171 (.003)	.131 (.004)	.110 (.003)	.197 (.003)	.172 (.005)	.201 (.004)
		FairRF	Workclass	.152 (.003)	.082 (.004)	.118 (.003)	.158 (.003)	.132 (.004)	.154 (.003)	.173 (.003)	.146 (.005)	.176 (.004)	.146 (.003)	.093 (.003)	.083 (.003)	.158 (.003)	.118 (.004)	.097 (.003)	.184 (.003)	.159 (.005)	.188 (.004)
	Race	CausalFT	Native-country	.163 (.003)	.130 (.003)	.089 (.003)	.176 (.003)	.157 (.004)	.106 (.003)	.189 (.003)	.199 (.004)	.186 (.004)	.152 (.003)	.123 (.003)	.088 (.003)	.174 (.003)	.147 (.004)	.116 (.003)	.194 (.003)	.221 (.005)	.208 (.004)
		FairRF	Education	.150 (.003)	.117 (.003)	.082 (.003)	.163 (.003)	.144 (.004)	.093 (.003)	.176 (.003)	.206 (.004)	.173 (.004)	.139 (.003)	.110 (.003)	.081 (.003)	.161 (.003)	.134 (.004)	.103 (.003)	.181 (.003)	.208 (.005)	.195 (.004)
	Age	CausalFT	Marital-status	.140 (.003)	.114 (.003)	.127 (.003)	.162 (.003)	.130 (.004)	.143 (.003)	.181 (.003)	.155 (.005)	.173 (.004)	.151 (.003)	.102 (.003)	.091 (.003)	.166 (.003)	.127 (.004)	.108 (.003)	.185 (.003)	.142 (.005)	.125 (.004)
		FairRF	Marital-status	.140 (.003)	.114 (.003)	.127 (.003)	.162 (.003)	.130 (.004)	.143 (.003)	.181 (.003)	.155 (.005)	.173 (.004)	.151 (.003)	.102 (.003)	.091 (.003)	.166 (.003)	.127 (.004)	.108 (.003)	.185 (.003)	.142 (.005)	.125 (.004)
Compas	Gender	CausalFT	Previous Convictions	.143 (.002)	.077 (.003)	.105 (.002)	.152 (.002)	.122 (.004)	.121 (.003)	.179 (.003)	.135 (.005)	.207 (.004)	.142 (.002)	.079 (.003)	.091 (.002)	.160 (.002)	.131 (.004)	.118 (.003)	.174 (.003)	.186 (.005)	.289 (.004)
		FairRF	Decile-text	.130 (.002)	.070 (.003)	.092 (.002)	.139 (.002)	.115 (.004)	.108 (.003)	.165 (.003)	.122 (.005)	.193 (.004)	.129 (.002)	.072 (.003)	.078 (.002)	.146 (.002)	.118 (.004)	.105 (.003)	.161 (.003)	.173 (.005)	.275 (.004)
	Race	CausalFT	Compas-score	.148 (.002)	.093 (.003)	.085 (.002)	.156 (.002)	.128 (.004)	.100 (.003)	.169 (.002)	.154 (.005)	.123 (.004)	.142 (.002)	.096 (.003)	.082 (.002)	.160 (.002)	.119 (.004)	.112 (.003)	.180 (.003)	.139 (.005)	.136 (.004)
		FairRF	Decile-text	.135 (.002)	.085 (.003)	.090 (.002)	.143 (.002)	.115 (.004)	.087 (.003)	.155 (.002)	.141 (.005)	.110 (.004)	.129 (.002)	.088 (.003)	.087 (.002)	.146 (.002)	.106 (.004)	.099 (.003)	.167 (.003)	.126 (.005)	.123 (.004)
Law School	Gender	CausalFT	Lsat	.153 (.003)	.108 (.004)	.072 (.002)	.163 (.003)	.125 (.003)	.109 (.003)	.175 (.003)	.152 (.004)	.206 (.004)	.141 (.003)	.110 (.003)	.077 (.002)	.161 (.003)	.146 (.004)	.116 (.003)	.171 (.003)	.199 (.005)	.251 (.004)
		FairRF	Decile1b	.142 (.004)	.098 (.005)	.065 (.003)	.152 (.004)	.115 (.004)	.099 (.004)	.164 (.004)	.142 (.005)	.195 (.005)	.130 (.004)	.100 (.004)	.078 (.003)	.150 (.004)	.136 (.005)	.106 (.004)	.160 (.004)	.189 (.006)	.240 (.005)
	Race	CausalFT	Lsat	.148 (.003)	.112 (.003)	.085 (.003)	.158 (.003)	.122 (.004)	.097 (.003)	.171 (.003)	.143 (.005)	.110 (.004)	.132 (.003)	.099 (.003)	.073 (.003)	.153 (.003)	.117 (.004)	.088 (.003)	.163 (.003)	.134 (.005)	.103 (.004)
		FairRF	Delice3	.138 (.004)	.102 (.004)	.075 (.004)	.148 (.004)	.112 (.005)	.087 (.004)	.161 (.004)	.133 (.006)	.100 (.005)	.122 (.004)	.103 (.004)	.063 (.004)	.143 (.004)	.107 (.005)	.078 (.004)	.153 (.004)	.124 (.006)	.093 (.005)
Kdd	Gender	CausalFT	Occupation	.148 (.003)	.099 (.003)	.130 (.003)	.162 (.003)	.133 (.004)	.148 (.004)	.170 (.003)	.189 (.005)	.235 (.004)	.138 (.003)	.106 (.003)	.106 (.003)	.157 (.003)	.143 (.004)	.135 (.003)	.177 (.003)	.249 (.005)	.172 (.004)
		FairRF	Housing	.135 (.004)	.088 (.004)	.118 (.004)	.149 (.004)	.120 (.005)	.135 (.005)	.157 (.004)	.176 (.006)	.222 (.005)	.125 (.004)	.093 (.004)	.094 (.004)	.144 (.004)	.130 (.005)	.122 (.004)	.164 (.004)	.236 (.006)	.159 (.005)
	Race	CausalFT	Tax-filer-stat	.152 (.003)	.146 (.003)	.171 (.003)	.164 (.003)	.155 (.004)	.179 (.003)	.173 (.003)	.198 (.005)	.250 (.004)	.145 (.003)	.151 (.003)	.162 (.003)	.161 (.003)	.184 (.004)	.202 (.003)	.181 (.003)	.253 (.005)	.209 (.004)
		FairRF	Marital-status	.139 (.004)	.133 (.004)	.158 (.004)	.151 (.004)	.142 (.005)	.166 (.004)	.160 (.004)	.185 (.006)	.237 (.005)	.132 (.004)	.138 (.004)	.149 (.004)	.148 (.004)	.171 (.005)	.189 (.004)	.168 (.004)	.240 (.006)	.196 (.005)
Dutch	Gender	CausalFT	Cur_eco_activity	.164 (.003)	.076 (.003)	.130 (.003)	.181 (.003)	.100 (.004)	.132 (.004)	.175 (.003)	.146 (.005)	.244 (.004)	.162 (.003)	.092 (.003)	.119 (.003)	.171 (.003)	.126 (.004)	.190 (.003)	.175 (.003)	.269 (.005)	.221 (.004)
		FairRF	Cur_eco_activity	.164 (.003)	.076 (.003)	.130 (.003)	.181 (.003)	.100 (.004)	.132 (.004)	.175 (.003)	.146 (.005)	.244 (.004)	.162 (.003)	.092 (.003)	.119 (.003)	.171 (.003)	.126 (.004)	.190 (.003)	.175 (.003)	.269 (.005)	.221 (.004)
	Age	CausalFT	Economic_status	.162 (.003)	.129 (.003)	.113 (.003)	.181 (.003)	.144 (.004)	.185 (.003)	.179 (.003)	.202 (.005)	.260 (.004)	.159 (.003)	.147 (.003)	.161 (.003)	.174 (.003)	.199 (.004)	.190 (.003)	.180 (.003)	.229 (.005)	.204 (.004)
		FairRF	Edu_level	.150 (.004)	.118 (.004)	.102 (.004)	.168 (.004)	.146 (.005)	.172 (.004)	.166 (.004)	.189 (.006)	.247 (.005)	.146 (.004)	.136 (.004)	.148 (.004)	.161 (.004)	.186 (.005)	.177 (.004)	.167 (.004)	.216 (.006)	.191 (.005)
Credit	Gender	CausalFT	Pay_amt3	.159 (.003)	.077 (.003)	.121 (.003)	.167 (.003)	.134 (.004)	.136 (.003)	.172 (.003)	.168 (.005)	.190 (.004)	.158 (.003)	.070 (.003)	.095 (.003)	.169 (.003)	.185 (.004)	.163 (.003)	.172 (.003)	.226 (.005)	.214 (.004)
		FairRF	Pay_amt3	.159 (.003)	.077 (.003)	.121 (.003)	.167 (.003)	.134 (.004)	.136 (.003)	.172 (.003)	.168 (.005)	.190 (.004)	.158 (.003)	.070 (.003)	.095 (.003)	.169 (.003)	.185 (.004)	.163 (.003)	.172 (.003)	.226 (.005)	.214 (.004)
	Marriage	CausalFT	Pay_amt1	.153 (.003)	.147 (.003)	.186 (.003)	.166 (.003)	.158 (.004)	.204 (.003)	.181 (.003)	.199 (.005)	.223 (.004)	.152 (.003)	.135 (.003)	.169 (.003)	.168 (.003)	.171 (.004)	.211 (.003)	.182 (.003)	.226 (.005)	.250 (.004)
		FairRF	Pay_amt1	.153 (.003)	.147 (.003)	.186 (.003)	.166 (.003)	.158 (.004)	.204 (.003)	.181 (.003)	.199 (.005)	.223 (.004)	.152 (.003)	.135 (.003)	.169 (.003)	.168 (.003)	.171 (.004)	.211 (.003)	.182 (.003)	.226 (.005)	.250 (.004)
	Education	CausalFT	Pay_amt1	.152 (.003)	.127 (.003)	.090 (.003)	.163 (.003)	.140 (.004)	.100 (.003)	.176 (.003)	.156 (.005)	.123 (.004)	.151 (.003)	.113 (.003)	.086 (.003)	.162 (.003)	.129 (.004)	.099 (.003)	.171 (.003)	.147 (.005)	.117 (.004)
		FairRF	BILL_AMT3	.140 (.004)	.115 (.004)	.078 (.004)	.150 (.004)	.128 (.005)	.088 (.004)	.163 (.004)	.143 (.006)	.110 (.005)	.138 (.004)	.101 (.004)	.074 (.004)	.149 (.004)	.131 (.005)	.087 (.004)	.158 (.004)	.134 (.006)	.104 (.005)
Crime	Race	CausalFT	Score_text	.159 (.003)	.112 (.003)	.093 (.003)	.180 (.003)	.137 (.004)	.113 (.003)	.191 (.003)	.174 (.005)	.258 (.004)	.163 (.003)	.102 (.003)	.090 (.003)	.171 (.003)	.152 (.004)	.130 (.003)	.190 (.003)	.218 (.005)	.276 (.004)
		FairRF	Score_text	.159 (.003)	.112 (.003)	.093 (.003)	.180 (.003)	.137 (.004)	.113 (.003)	.191 (.003)	.174 (.005)	.258 (.004)	.163 (.003)	.102 (.003)	.090 (.003)	.171 (.003)	.152 (.004)	.130 (.003)	.190 (.003)	.218 (.005)	.276 (.004)
	Gender	CausalFT	Priors_count	.168 (.003)	.073 (.003)	.089 (.003)	.186 (.003)	.123 (.004)	.135 (.003)	.195 (.003)	.202 (.005)	.203 (.004)	.166 (.003)	.092 (.003)	.104 (.003)	.180 (.003)	.140 (.004)	.178 (.003)	.195 (.003)	.223 (.005)	.214 (.004)
		FairRF	V_score_text	.155 (.004)	.061 (.004)	.077 (.004)	.173 (.004)	.111 (.005)	.123 (.004)	.182 (.004)	.189 (.006)	.190 (.005)	.153 (.004)	.080 (.004)	.092 (.004)	.167 (.004)	.128 (.005)	.165 (.004)	.182 (.004)	.210 (.006)	.201 (.005)
German	Gender	CausalFT	Credit-amount	.167 (.003)	.133 (.003)	.106 (.003)	.181 (.003)	.153 (.004)	.129 (.003)	.188 (.003)	.223 (.005)	.189 (.004)	.165 (.003)	.117 (.003)	.096 (.003)	.174 (.003)	.173 (.004)	.131 (.003)	.186 (.003)	.238 (.005)	.231 (.004)
		FairRF	Credit-amount	.167 (.003)	.133 (.003)	.106 (.003)	.181 (.003)	.153 (.004)	.129 (.003)	.188 (.003)	.223 (.005)	.189 (.004)	.165 (.003)	.117 (.003)	.096 (.003)	.174 (.003)	.173 (.004)	.131 (.003)	.186 (.003)	.238 (.005)	.231 (.004)
	Age	CausalFT	Existing-credits	.173 (.003)	.107 (.003)	.139 (.003)	.186 (.003)	.183 (.004)	.170 (.003)	.188 (.003)	.218 (.005)	.189 (.004)	.171 (.003)	.130 (.003)	.132 (.003)	.177 (.003)	.196 (.004)	.172 (.003)	.190 (.003)	.260 (.005)	.215 (.004)
		FairRF	Duration	.160 (.004)	.095 (.004)	.127 (.004)	.173 (.004)	.170 (.005)	.157 (.004)	.175 (.004)	.205 (.006)	.176 (.005)	.158 (.004)	.118 (.004)	.120 (.004)	.164 (.004)	.183 (.005)	.159 (.004)	.177 (.004)	.247 (.006)	.202 (.005)

- **I:** The number of unique invalid pairs within those generated by CausalFT before invalidity repair.
- **IR:** The number of unique pairs that have been successfully repaired from those invalid ones generated by CausalFT before invalidity repair.

5.3.2 Results. As shown in Table 10, for all cases, the number of pairs generated with CausalFT is much higher than the case when directly using the base generator. This makes sense, since the perturbation in CausalFT with relaxed individual discriminatory instance definition helps to increase the chance of successfully pairing with the original samples in the testing data (even under the true individual discriminatory instance definition), together with the possibility of creating invalid samples. Within those generated pairs, CausalFT produces around ≈ 24% invalid ones, which can be successfully repaired by around ≈ 81% on average for all cases. This demonstrates the effectiveness of the invalidity repairs that can significantly improve the ability to find fairness bugs. As a result, we conclude that:

Table 10. Counting the average number of pairs (test cases) generated with the relaxed individual discriminatory instance definition and invalidity repair across all base generators and runs.

Dataset	f_s	DNN$_5$				DNN$_6$				LR				RF			
		#w/o R	#w/ R	#I	#IR	#w/o R	#w/ R	#I	#IR	#w/o R	#w/ R	#I	#IR	#w/o R	#w/ R	#I	#IR
Adult	*Gender*	2614.9	4314.3	649.0	475.9	2811.7	4589.7	683.2	511.7	1738.3	2898.3	554.7	449.3	1609.3	2795.5	645.1	522.4
	Race	2991.5	5301.3	893.0	721.5	3216.7	5639.7	940.0	775.8	1680.0	2622.3	400.0	336.0	1635.0	2946.8	679.3	509.5
	Age	2816.3	4877.0	940.2	713.3	3028.3	5188.3	989.7	767.0	1566.7	2776.5	573.0	446.9	1531.3	2677.0	572.6	458.1
Compas	*Gender*	3072.9	5099.2	775.9	569.6	3304.2	5424.7	816.7	612.5	1618.3	2662.0	462.0	378.8	1521.0	2526.8	541.6	442.6
	Race	3377.5	5483.4	1052.7	814.1	3631.7	5833.4	1108.1	875.4	1578.3	2918.8	583.8	472.9	1464.0	2416.2	484.0	401.7
Law School	*Gender*	2823.3	4674.6	728.5	570.5	3035.8	4973.0	766.8	613.4	1553.3	2790.5	627.3	483.0	1559.5	2467.4	502.4	414.2
	Race	2814.8	5006.7	918.7	764.5	3026.7	5326.3	967.1	822.0	1561.7	2703.6	431.7	345.4	1499.5	2397.4	483.8	398.1
KDD	*Gender*	2359.1	4191.6	610.7	495.5	2536.7	4459.1	642.8	532.8	1551.7	2523.2	352.5	299.6	1555.0	2530.5	506.1	415.0
	Race	2366.8	4283.7	721.5	561.8	2545.0	4557.1	759.5	604.1	1546.7	2937.1	529.6	418.4	1578.2	2638.2	530.0	434.6
Dutch	*Gender*	2340.5	3716.8	485.3	380.0	2516.7	3954.0	510.8	408.6	1520.0	2661.2	476.7	384.9	1653.3	2618.1	581.2	476.6
	Age	2315.7	3666.9	546.3	438.6	2490.0	3901.0	575.1	471.6	1577.0	2821.0	528.8	444.2	1683.8	2702.0	624.6	502.3
Credit	*Gender*	2321.9	3449.2	528.8	429.1	2496.7	3669.4	556.6	461.4	1577.0	2603.8	491.4	408.2	1596.2	2560.0	546.8	437.4
	Marriage	2604.0	4157.9	720.7	545.0	2800.0	4423.3	758.6	586.0	1517.0	2816.5	567.2	465.1	1622.2	2624.2	579.0	486.4
	Education	2456.8	3976.2	687.1	538.1	2641.7	4230.0	723.3	578.6	1740.0	3046.2	608.7	492.3	1574.2	2500.6	513.8	410.1
Crime	*Race*	2332.0	3690.5	562.1	440.3	2507.5	3926.1	591.7	473.4	1472.0	2485.2	461.4	372.7	1717.0	2726.4	592.3	503.3
	Gender	2573.0	4241.1	776.7	608.3	2766.7	4511.8	817.6	654.1	1712.0	3017.0	607.2	516.1	1766.7	2811.8	602.0	507.7
German	*Gender*	2439.7	3952.7	704.4	544.8	2623.3	4205.0	741.5	585.8	1535.2	2762.2	570.7	468.0	1718.5	2694.6	536.1	455.7
	Age	2433.5	3908.6	661.2	517.8	2616.7	4158.1	696.0	556.8	1536.4	2631.3	530.3	444.7	1755.8	2765.4	567.2	479.1

> ***RQ$_3$:** The relaxed IDI definition in* `CausalFT` *can naturally produce more pairs of test samples, within which the invalid ones can be effectively repaired to boost its ability in revealing fairness issues.*

5.4 Fairness Improvement

5.4.1 Method. Although this work focuses on testing, in **RQ$_4$**, we verify how the samples generated by `CausalFT`, after correcting the pair to produce the same outcome, can improve the AI models in better-mitigating fairness bugs, as in prior work [79].

To that end, for each dataset, we compare `CausalFT` with a base generator without `CausalFT` and `FairRF`, all of which are under their corresponding best generators, in the steps below (for all cases):

(1) Run a testing approach on the AI models to generate test cases (pairs) that reveal the fairness bugs.
(2) Repair the fairness bugs found, e.g., if $\mathbf{x'}_a$ and $\mathbf{x'}_b$ lead to different labels in the prediction and $\mathbf{x'}_a$ is from the original testing data, we then correct them by changing $\mathbf{x'}_b$ to have the same label as $\mathbf{x'}_a$.
(3) Feed the corrected pairs of samples to retrain the AI model.
(4) The improved model is then tested again by using `CausalFT` under its best generator again for 10 runs.

We also compare with the testing results for the AI model before retraining, denoted as `Before`. If the samples generated by `CausalFT` can better improve the AI model in handling fairness, the improved DNN should be harder to test, leading to a lower IDI ratio, EOD, and SPD value. In addition, we have also reported on the changes in the model quality via accuracy, F1-score, and AUC, since the fairness can be conflicting with those metrics; though it is worth noting that the model quality might not imply practical usefulness [18].

5.4.2 Result. For the impact on model quality, as can be seen from Tables 11, 12, 13, and 14, we see that there is indeed a general degradation in the model quality in terms of accuracy and AUC. This makes sense, as it is known that the fairness could conflict with the model quality [13, 49, 73]. However, in general, we see that the degradation is relatively smaller compared with the improvement observed for fairness preservation, e.g., when compared with `Before` on DNN$_6$,

Table 11. Testing the retrained DNN_5 improved by CausalFT against before retraining, without CausalFT, and FairRF over 5 retraining runs × 10 testing runs (i.e., $A - B$ where A and B are the metric values of testing the retrained model by other and CausalFT, respectively). All retrained samples are generated with the best generator under an approach. Positive numbers imply CausalFT improves more. green cells mean the improvement is statistically significant. red cells mean the decrease is statistically significant.

Dataset	f_s	IDI ratio (CausalFT vs.)			EOD (CausalFT vs.)			SPD (CausalFT vs.)			ACC (CausalFT vs.)			F1 (CausalFT vs.)			AUC (CausalFT vs.)		
		Before	w/o	FairRF	Before	w/o	FairRF	Before	w/o	FairRF	Before	w/o	FairRF	Before	w/o	FairRF	Before	w/o	FairRF
Adult	Gender	.219(.002)	.035(.003)	.044(.003)	.086(.002)	.018(.002)	.022(.002)	.043(.002)	.011(.002)	.014(.002)	-.050(.002)	-.027(.002)	-.030(.002)	.032(.002)	.012(.002)	.015(.002)	-.027(.002)	-.010(.002)	-.011(.002)
	Race	.244(.003)	.040(.003)	.043(.003)	.080(.002)	.022(.002)	.028(.002)	.072(.002)	.013(.002)	.016(.002)	-.053(.002)	-.026(.002)	-.026(.002)	.036(.002)	.014(.002)	.014(.002)	-.033(.002)	-.012(.002)	-.010(.002)
	Age	.229(.003)	.035(.003)	.023(.005)	.089(.002)	.017(.002)	.014(.004)	.064(.002)	.009(.001)	.012(.004)	-.050(.002)	-.021(.002)	-.023(.002)	.030(.002)	.013(.002)	.015(.002)	-.036(.002)	-.010(.002)	-.009(.002)
Compas	Gender	.254(.003)	.039(.003)	.046(.004)	.082(.002)	.019(.002)	.025(.002)	.050(.002)	.013(.002)	.012(.002)	-.052(.002)	-.022(.002)	-.027(.002)	.031(.002)	.012(.002)	.013(.002)	-.026(.002)	-.009(.002)	-.010(.002)
	Race	.292(.003)	.034(.004)	.050(.003)	.094(.002)	.022(.003)	.024(.002)	.079(.002)	.011(.002)	.010(.001)	-.049(.002)	-.020(.002)	-.020(.002)	.034(.002)	.017(.002)	.016(.002)	-.030(.002)	-.010(.002)	-.011(.002)
Law School	Gender	.233(.003)	.035(.003)	.043(.003)	.089(.002)	.018(.004)	.019(.002)	.068(.002)	.012(.002)	.008(.001)	-.044(.002)	-.019(.002)	-.022(.002)	.027(.002)	.013(.002)	.017(.002)	-.021(.002)	-.010(.002)	-.012(.002)
	Race	.230(.003)	.040(.003)	.039(.003)	.078(.003)	.023(.003)	.022(.002)	.062(.002)	.014(.002)	.010(.001)	-.046(.002)	-.020(.002)	-.019(.002)	.032(.002)	.015(.002)	.016(.002)	-.028(.002)	-.018(.002)	-.017(.002)
Kdd	Gender	.186(.003)	.028(.003)	.036(.002)	.083(.002)	.017(.003)	.020(.003)	.060(.002)	.010(.002)	.012(.004)	-.043(.002)	-.018(.002)	-.017(.002)	.026(.002)	.012(.002)	.011(.002)	-.018(.002)	-.008(.002)	-.006(.002)
	Race	.194(.003)	.033(.004)	.051(.003)	.074(.002)	.019(.003)	.018(.002)	.065(.002)	.012(.004)	.011(.001)	-.039(.002)	-.017(.002)	-.016(.002)	.024(.002)	.011(.002)	.011(.002)	-.021(.002)	-.009(.002)	-.008(.002)
Dutch	Gender	.206(.003)	.020(.003)	.020(.007)	.078(.002)	.009(.004)	.010(.004)	.070(.002)	.004(.007)	.006(.006)	-.047(.002)	-.021(.002)	-.020(.002)	.030(.002)	.015(.002)	.012(.002)	-.019(.002)	-.011(.002)	-.009(.002)
	Age	.191(.003)	.023(.003)	.052(.004)	.084(.002)	.017(.002)	.022(.002)	.076(.002)	.009(.003)	.010(.001)	-.049(.002)	-.019(.002)	-.020(.002)	.034(.002)	.011(.002)	.014(.002)	-.022(.002)	-.012(.002)	-.011(.002)
Credit	Gender	.181(.003)	.032(.003)	.022(.005)	.090(.002)	.013(.002)	.015(.004)	.068(.002)	.009(.001)	.010(.003)	-.057(.002)	-.021(.002)	-.023(.002)	.028(.002)	.015(.002)	.014(.002)	-.020(.002)	-.012(.002)	-.008(.002)
	Marriage	.183(.003)	.027(.003)	.019(.004)	.085(.002)	.011(.005)	.014(.004)	.063(.002)	.008(.002)	.010(.006)	-.050(.002)	-.017(.002)	-.013(.002)	.031(.002)	.012(.002)	.016(.002)	-.028(.002)	-.010(.002)	-.011(.002)
	Education	.187(.003)	.033(.003)	.041(.003)	.082(.002)	.011(.004)	.013(.002)	.067(.002)	.011(.003)	.007(.001)	-.048(.002)	-.022(.002)	-.017(.002)	.025(.002)	.012(.002)	.013(.002)	-.023(.002)	-.013(.002)	-.014(.002)
Crime	Race	.179(.003)	.030(.003)	.029(.006)	.088(.002)	.007(.004)	.012(.004)	.064(.002)	.009(.002)	.007(.003)	-.048(.002)	-.017(.002)	-.017(.002)	.031(.002)	.014(.002)	.015(.002)	-.030(.002)	-.010(.002)	-.012(.002)
	Gender	.187(.003)	.024(.005)	.040(.003)	.085(.002)	.010(.003)	.016(.002)	.070(.002)	.008(.003)	.011(.001)	-.046(.002)	-.020(.002)	-.015(.002)	.032(.002)	.013(.002)	.016(.002)	-.026(.002)	-.009(.002)	-.014(.002)
German	Gender	.173(.003)	.040(.002)	.027(.006)	.078(.002)	.016(.002)	.012(.002)	.075(.002)	.012(.002)	.004(.001)	-.038(.002)	-.019(.002)	-.018(.002)	.021(.002)	.011(.002)	.009(.002)	-.020(.002)	-.012(.002)	-.015(.002)
	Age	.175(.003)	.043(.003)	.033(.003)	.083(.002)	.015(.002)	.018(.002)	.067(.002)	.009(.001)	.010(.002)	-.049(.002)	-.019(.002)	-.018(.002)	.033(.002)	.016(.002)	.017(.002)	-.031(.001)	-.018(.002)	-.017(.002)

Table 12. Testing the retrained DNN_6 model improved by CausalFT against before retraining, without CausalFT, and FairRF over 5 retraining runs × 10 testing runs. Other formate is the same as Table 11.

Dataset	f_s	IDI ratio (CausalFT vs.)			EOD (CausalFT vs.)			SPD (CausalFT vs.)			ACC (CausalFT vs.)			F1 (CausalFT vs.)			AUC (CausalFT vs.)		
		Before	w/o	FairRF	Before	w/o	FairRF	Before	w/o	FairRF	Before	w/o	FairRF	Before	w/o	FairRF	Before	w/o	FairRF
Adult	Gender	.214(.002)	.033(.003)	.046(.003)	.084(.002)	.017(.002)	.021(.002)	.041(.002)	.012(.002)	.013(.002)	-.052(.002)	-.029(.002)	-.031(.002)	.033(.002)	.012(.002)	.014(.002)	-.028(.002)	-.011(.002)	-.010(.002)
	Race	.248(.003)	.042(.003)	.041(.003)	.078(.002)	.023(.002)	.029(.002)	.074(.002)	.014(.002)	.017(.002)	-.055(.002)	-.028(.002)	-.027(.002)	.038(.002)	.015(.002)	.014(.002)	-.035(.002)	-.013(.002)	-.011(.002)
	Age	.225(.003)	.036(.003)	.021(.006)	.091(.002)	.016(.002)	.013(.005)	.066(.002)	.008(.001)	.011(.005)	-.051(.002)	-.022(.002)	-.024(.002)	.031(.002)	.013(.002)	.015(.002)	-.037(.002)	-.011(.002)	-.008(.002)
Compas	Gender	.257(.003)	.038(.003)	.044(.004)	.081(.002)	.018(.002)	.026(.002)	.051(.002)	.014(.002)	.011(.002)	-.054(.002)	-.023(.002)	-.028(.002)	.032(.002)	.012(.002)	.013(.002)	-.027(.002)	-.008(.002)	-.011(.002)
	Race	.290(.003)	.036(.004)	.051(.003)	.092(.002)	.021(.003)	.023(.002)	.081(.002)	.012(.002)	.010(.001)	-.051(.002)	-.021(.002)	-.020(.002)	.035(.002)	.018(.002)	.017(.002)	-.031(.002)	-.010(.002)	-.011(.002)
Law School	Gender	.235(.003)	.036(.003)	.042(.003)	.088(.002)	.019(.005)	.019(.001)	.069(.002)	.013(.002)	.008(.001)	-.045(.002)	-.020(.002)	-.022(.002)	.026(.002)	.013(.002)	.017(.002)	-.022(.002)	-.011(.002)	-.013(.002)
	Race	.228(.003)	.039(.003)	.041(.003)	.077(.003)	.024(.003)	.023(.002)	.063(.002)	.015(.002)	.011(.001)	-.048(.002)	-.021(.002)	-.020(.002)	.033(.002)	.018(.002)	.017(.002)	-.029(.002)	-.020(.002)	-.018(.002)
Kdd	Gender	.184(.003)	.027(.003)	.037(.002)	.082(.002)	.016(.003)	.021(.003)	.061(.002)	.010(.002)	.012(.004)	-.044(.002)	-.019(.002)	-.018(.002)	.027(.002)	.012(.002)	.011(.002)	-.019(.002)	-.009(.002)	-.006(.002)
	Race	.192(.003)	.034(.005)	.052(.003)	.073(.002)	.020(.003)	.017(.002)	.064(.002)	.013(.004)	.011(.001)	-.040(.002)	-.018(.002)	-.016(.002)	.025(.002)	.012(.002)	.011(.002)	-.022(.002)	-.009(.002)	-.008(.002)
Dutch	Gender	.205(.003)	.021(.003)	.019(.008)	.079(.002)	.008(.004)	.009(.005)	.071(.002)	.004(.008)	.007(.006)	-.048(.002)	-.022(.002)	-.020(.002)	.031(.002)	.016(.002)	.012(.002)	-.020(.002)	-.012(.002)	-.009(.002)
	Age	.189(.003)	.024(.003)	.054(.004)	.083(.003)	.018(.002)	.023(.002)	.077(.002)	.009(.003)	.011(.001)	-.050(.002)	-.020(.002)	-.021(.002)	.036(.002)	.011(.002)	.014(.002)	-.023(.002)	-.013(.002)	-.011(.002)
Credit	Gender	.179(.003)	.033(.003)	.021(.006)	.091(.002)	.014(.002)	.016(.005)	.069(.002)	.010(.001)	.009(.004)	-.058(.002)	-.022(.002)	-.024(.002)	.029(.002)	.016(.002)	.015(.002)	-.021(.002)	-.013(.002)	-.007(.002)
	Marriage	.182(.003)	.028(.003)	.018(.005)	.086(.002)	.011(.006)	.013(.004)	.064(.002)	.008(.002)	.010(.007)	-.052(.002)	-.018(.002)	-.013(.002)	.032(.002)	.012(.002)	.017(.002)	-.027(.002)	-.011(.002)	-.012(.002)
	Education	.186(.003)	.034(.003)	.042(.003)	.083(.002)	.012(.004)	.014(.001)	.068(.002)	.011(.003)	.006(.001)	-.050(.002)	-.023(.002)	-.019(.002)	.034(.002)	.011(.002)	.013(.002)	-.032(.002)	-.010(.002)	-.010(.002)
Credit	Gender	.179(.003)	.033(.003)	.021(.006)	.091(.002)	.014(.002)	.016(.005)	.069(.002)	.010(.001)	.009(.004)	-.048(.002)	-.019(.002)	-.017(.002)	.033(.002)	.010(.002)	.010(.002)	-.031(.002)	-.011(.002)	-.010(.002)
	Marriage	.182(.003)	.028(.003)	.018(.005)	.086(.002)	.011(.006)	.013(.004)	.064(.002)	.008(.002)	.010(.007)	-.039(.002)	-.015(.002)	-.018(.002)	.022(.002)	.006(.002)	.017(.002)	-.030(.002)	-.007(.002)	-.011(.002)
	Education	.186(.003)	.034(.003)	.042(.003)	.083(.002)	.012(.004)	.014(.001)	.068(.002)	.011(.003)	.006(.001)	-.041(.002)	-.020(.002)	-.019(.002)	.024(.002)	.012(.002)	.013(.002)	-.022(.002)	-.014(.002)	-.015(.002)
Crime	Race	.178(.003)	.031(.003)	.028(.007)	.089(.002)	.006(.005)	.013(.005)	.063(.002)	.010(.002)	.007(.003)	-.049(.002)	-.018(.002)	-.017(.002)	.032(.002)	.015(.002)	.016(.002)	-.031(.002)	-.011(.002)	-.012(.002)
	Gender	.186(.003)	.025(.006)	.041(.003)	.086(.002)	.011(.003)	.017(.001)	.071(.002)	.008(.003)	.011(.001)	-.047(.002)	-.021(.002)	-.015(.002)	.033(.002)	.013(.002)	.017(.002)	-.027(.002)	-.009(.002)	-.015(.002)
German	Gender	.171(.003)	.041(.002)	.026(.008)	.079(.002)	.017(.002)	.012(.001)	.076(.002)	.013(.002)	.004(.001)	-.039(.002)	-.020(.002)	-.018(.002)	.022(.002)	.011(.002)	.009(.001)	-.021(.002)	-.013(.002)	-.016(.002)
	Age	.174(.003)	.044(.003)	.032(.003)	.084(.002)	.016(.002)	.019(.002)	.068(.002)	.009(.001)	.010(.002)	-.050(.002)	-.020(.002)	-.019(.002)	.034(.002)	.017(.002)	.018(.002)	-.032(.001)	-.019(.002)	-.017(.002)

Table 13. Testing the retrained LR model improved by CausalFT against before retraining, without CausalFT, and FairRF over 5 retraining runs × 10 testing runs. Other formate is the same as Table 11.

Dataset	f_s	IDI ratio (CausalFT vs.)			EOD (CausalFT vs.)			SPD (CausalFT vs.)			ACC (CausalFT vs.)			F1 (CausalFT vs.)			AUC (CausalFT vs.)		
		Before	w/o	FairRF	Before	w/o	FairRF	Before	w/o	FairRF	Before	w/o	FairRF	Before	w/o	FairRF	Before	w/o	FairRF
Adult	Gender	.081(.003)	.065(.003)	.072(.003)	.028(.002)	.023(.002)	.026(.002)	.019(.002)	.015(.002)	.017(.002)	-.013(.002)	-.010(.002)	-.012(.002)	.014(.002)	.011(.002)	.013(.002)	.013(.002)	.010(.002)	.012(.002)
	Race	.087(.003)	.070(.003)	.075(.003)	.030(.002)	.025(.002)	.028(.002)	.020(.002)	.016(.002)	.018(.002)	-.014(.002)	-.011(.002)	-.013(.002)	.015(.002)	.012(.002)	.014(.002)	.014(.002)	.011(.002)	.013(.002)
	Age	.085(.003)	.068(.003)	.073(.003)	.029(.002)	.024(.002)	.027(.002)	.019(.002)	.015(.002)	.017(.002)	-.013(.002)	-.010(.002)	-.012(.002)	.014(.002)	.011(.002)	.013(.002)	.013(.002)	.010(.002)	.012(.002)
Compas	Gender	.079(.003)	.063(.003)	.069(.003)	.027(.002)	.022(.002)	.025(.002)	.018(.002)	.015(.002)	.016(.002)	-.012(.002)	-.010(.002)	-.011(.002)	.013(.002)	.011(.002)	.012(.002)	.012(.002)	.010(.002)	.011(.002)
	Race	.084(.003)	.068(.003)	.073(.003)	.028(.002)	.023(.002)	.026(.002)	.019(.002)	.016(.002)	.017(.002)	-.013(.002)	-.010(.002)	-.012(.002)	.014(.002)	.011(.002)	.013(.002)	.013(.002)	.010(.002)	.012(.002)
Law School	Gender	.077(.003)	.061(.003)	.066(.003)	.026(.002)	.021(.002)	.024(.002)	.018(.002)	.014(.002)	.016(.002)	-.012(.002)	-.009(.002)	-.011(.002)	.013(.002)	.010(.002)	.012(.002)	.012(.002)	.009(.002)	.011(.002)
	Race	.080(.003)	.063(.003)	.068(.003)	.027(.002)	.022(.002)	.025(.002)	.018(.002)	.015(.002)	.017(.002)	-.013(.002)	-.010(.002)	-.012(.002)	.014(.002)	.011(.002)	.013(.002)	.013(.002)	.010(.002)	.012(.002)
Kdd	Gender	.078(.003)	.062(.003)	.067(.003)	.026(.002)	.021(.002)	.024(.002)	.018(.002)	.015(.002)	.016(.002)	-.012(.002)	-.009(.002)	-.011(.002)	.013(.002)	.010(.002)	.012(.002)	.012(.002)	.009(.002)	.011(.002)
	Race	.082(.003)	.065(.003)	.070(.003)	.028(.002)	.023(.002)	.025(.002)	.019(.002)	.016(.002)	.017(.002)	-.013(.002)	-.010(.002)	-.012(.002)	.014(.002)	.011(.002)	.013(.002)	.013(.002)	.010(.002)	.012(.002)
Dutch	Gender	.080(.003)	.063(.003)	.068(.003)	.027(.002)	.022(.002)	.025(.002)	.018(.002)	.015(.002)	.016(.002)	-.013(.002)	-.010(.002)	-.012(.002)	.014(.002)	.011(.002)	.013(.002)	.013(.002)	.010(.002)	.012(.002)
	Age	.078(.003)	.061(.003)	.066(.003)	.026(.002)	.021(.002)	.024(.002)	.017(.002)	.014(.002)	.015(.002)	-.012(.002)	-.009(.002)	-.011(.002)	.013(.002)	.010(.002)	.012(.002)	.012(.002)	.009(.002)	.011(.002)
Credit	Gender	.079(.003)	.062(.003)	.067(.003)	.027(.002)	.021(.002)	.024(.002)	.018(.002)	.015(.002)	.016(.002)	-.012(.002)	-.009(.002)	-.011(.002)	.013(.002)	.010(.002)	.012(.002)	.012(.002)	.009(.002)	.011(.002)
	Marriage	.081(.003)	.065(.003)	.070(.003)	.028(.002)	.023(.002)	.026(.002)	.019(.002)	.016(.002)	.017(.002)	-.013(.002)	-.010(.002)	-.012(.002)	.014(.002)	.011(.002)	.013(.002)	.013(.002)	.010(.002)	.012(.002)
	Education	.080(.003)	.064(.003)	.069(.003)	.027(.002)	.022(.002)	.025(.002)	.018(.002)	.015(.002)	.016(.002)	-.013(.002)	-.010(.002)	-.012(.002)	.014(.002)	.011(.002)	.013(.002)	.013(.002)	.010(.002)	.012(.002)
Crime	Race	.077(.003)	.061(.003)	.066(.003)	.026(.002)	.021(.002)	.024(.002)	.018(.002)	.014(.002)	.016(.002)	-.012(.002)	-.009(.002)	-.011(.002)	.013(.002)	.010(.002)	.012(.002)	.012(.002)	.009(.002)	.011(.002)
	Gender	.079(.003)	.063(.003)	.068(.003)	.027(.002)	.022(.002)	.025(.002)	.018(.002)	.015(.002)	.016(.002)	-.013(.002)	-.010(.002)	-.012(.002)	.014(.002)	.011(.002)	.013(.002)	.013(.002)	.010(.002)	.012(.002)
German	Gender	.078(.003)	.061(.003)	.066(.003)	.026(.002)	.021(.002)	.024(.002)	.018(.002)	.014(.002)	.016(.002)	-.012(.002)	-.009(.002)	-.011(.002)	.013(.002)	.010(.002)	.012(.002)	.012(.002)	.009(.002)	.011(.002)
	Age	.080(.003)	.063(.003)	.068(.003)	.027(.002)	.022(.002)	.025(.002)	.018(.002)	.015(.002)	.016(.002)	-.013(.002)	-.010(.002)	-.012(.002)	.014(.002)	.011(.002)	.013(.002)	.013(.002)	.010(.002)	.012(.002)

the biggest degradations on accuracy and AUC are 0.058 and 0.052, respectively, but in return, there

Table 14. Testing the retrained RF model improved by CausalFT against before retraining, without CausalFT, and FairRF over 5 retraining runs × 10 testing runs. Other formate is the same as Table 11.

Dataset	f_s	IDI ratio (CausalFT vs.)			EOD (CausalFT vs.)			SPD (CausalFT vs.)			ACC (CausalFT vs.)			F1 (CausalFT vs.)			AUC (CausalFT vs.)		
		Before	w/o	FairRF	Before	w/o	FairRF	Before	w/o	FairRF	Before	w/o	FairRF	Before	w/o	FairRF	Before	w/o	FairRF
Adult	*Gender*	.067(.003)	.018(.002)	.022(.002)	.027(.002)	.009(.002)	.010(.002)	.021(.002)	.006(.002)	.007(.002)	-.011(.002)	-.005(.001)	-.006(.002)	.013(.002)	.006(.002)	.007(.002)	-.010(.002)	-.005(.001)	-.006(.002)
	Race	.073(.003)	.020(.002)	.018(.002)	.023(.002)	.010(.002)	.012(.002)	.025(.002)	.007(.002)	.008(.002)	-.012(.002)	-.006(.002)	-.007(.002)	.015(.002)	.007(.002)	.008(.002)	-.012(.002)	-.006(.002)	-.007(.002)
	Age	.069(.003)	.019(.002)	.016(.002)	.030(.003)	.008(.002)	.009(.002)	.023(.002)	.005(.001)	.007(.002)	-.010(.002)	-.004(.001)	-.005(.001)	.012(.002)	.005(.001)	.006(.002)	-.011(.002)	-.005(.001)	-.006(.002)
Compas	*Gender*	.080(.003)	.021(.002)	.025(.002)	.028(.002)	.010(.002)	.012(.002)	.020(.002)	.007(.002)	.009(.002)	-.011(.002)	-.005(.001)	-.006(.002)	.013(.002)	.006(.002)	.007(.002)	-.010(.002)	-.005(.001)	-.006(.002)
	Race	.083(.003)	.020(.002)	.023(.002)	.031(.003)	.011(.002)	.013(.002)	.024(.002)	.008(.002)	.009(.002)	-.012(.002)	-.006(.002)	-.007(.002)	.014(.002)	.007(.002)	.008(.002)	-.012(.002)	-.006(.002)	-.007(.002)
Law School	*Gender*	.062(.003)	.018(.002)	.019(.002)	.025(.002)	.009(.002)	.010(.002)	.019(.002)	.007(.002)	.008(.002)	-.009(.002)	-.005(.001)	-.006(.002)	.011(.002)	.005(.001)	.006(.002)	-.009(.002)	-.005(.002)	-.006(.002)
	Race	.065(.003)	.020(.002)	.021(.002)	.027(.002)	.010(.002)	.011(.002)	.021(.002)	.007(.002)	.008(.002)	-.010(.002)	-.005(.002)	-.006(.002)	.012(.002)	.006(.002)	.007(.002)	-.010(.002)	-.005(.001)	-.006(.002)
Kdd	*Gender*	.071(.003)	.019(.002)	.023(.002)	.026(.002)	.009(.002)	.011(.002)	.020(.002)	.009(.002)	.008(.002)	-.010(.002)	-.005(.001)	-.006(.001)	.012(.002)	.006(.002)	.007(.002)	-.010(.002)	-.005(.001)	-.006(.002)
	Race	.075(.003)	.020(.002)	.022(.002)	.029(.002)	.010(.002)	.012(.002)	.022(.002)	.007(.002)	.008(.002)	-.011(.002)	-.005(.001)	-.006(.002)	.013(.002)	.006(.002)	.007(.002)	-.011(.002)	-.005(.001)	-.006(.002)
Dutch	*Gender*	.064(.003)	.017(.002)	.019(.002)	.026(.002)	.008(.002)	.009(.002)	.020(.002)	.006(.001)	.007(.002)	-.010(.002)	-.005(.001)	-.006(.002)	.012(.002)	.006(.002)	.007(.002)	-.010(.002)	-.005(.001)	-.006(.002)
	Age	.060(.003)	.018(.002)	.020(.002)	.027(.002)	.009(.002)	.010(.002)	.021(.002)	.007(.002)	.008(.002)	-.011(.002)	-.005(.001)	-.006(.002)	.013(.002)	.006(.002)	.007(.002)	-.011(.002)	-.005(.001)	-.006(.002)
Credit	*Gender*	.058(.003)	.015(.002)	.017(.002)	.025(.002)	.008(.002)	.009(.002)	.019(.002)	.006(.001)	.007(.002)	-.009(.002)	-.004(.001)	-.005(.001)	.011(.002)	.005(.001)	.006(.002)	-.009(.002)	-.005(.001)	-.006(.002)
	Marriage	.061(.003)	.016(.002)	.018(.002)	.024(.002)	.007(.002)	.009(.002)	.020(.002)	.006(.001)	.007(.002)	-.010(.002)	-.005(.001)	-.006(.002)	.012(.002)	.006(.002)	.007(.002)	-.010(.002)	-.005(.001)	-.006(.002)
	Education	.065(.003)	.017(.002)	.019(.002)	.026(.002)	.008(.002)	.009(.002)	.021(.002)	.007(.002)	.008(.002)	-.011(.002)	-.005(.001)	-.006(.002)	.013(.002)	.006(.002)	.007(.002)	-.011(.002)	-.005(.001)	-.006(.002)
Crime	*Race*	.056(.003)	.015(.002)	.017(.002)	.023(.002)	.007(.002)	.008(.002)	.018(.002)	.005(.001)	.006(.002)	-.009(.002)	-.004(.001)	-.005(.001)	.011(.002)	.005(.001)	.006(.002)	-.009(.002)	-.005(.001)	-.006(.002)
	Gender	.059(.003)	.016(.002)	.018(.002)	.024(.002)	.008(.002)	.009(.002)	.019(.002)	.006(.001)	.007(.002)	-.010(.002)	-.005(.001)	-.006(.002)	.012(.002)	.006(.002)	.007(.002)	-.010(.002)	-.005(.001)	-.006(.002)
German	*Gender*	.053(.003)	.014(.002)	.016(.002)	.022(.002)	.007(.002)	.008(.002)	.017(.002)	.005(.001)	.006(.002)	-.009(.002)	-.004(.001)	-.005(.001)	.011(.002)	.005(.001)	.006(.002)	-.009(.002)	-.005(.001)	-.006(.002)
	Age	.055(.003)	.015(.002)	.017(.002)	.023(.002)	.008(.002)	.009(.002)	.018(.002)	.006(.001)	.007(.002)	-.009(.002)	-.004(.001)	-.005(.001)	.011(.002)	.005(.001)	.006(.002)	-.009(.002)	-.005(.001)	-.006(.002)

is up to 0.290 improvement on IDI ratio. Interestingly, we also observe an improvement on the F1-score (and sometimes the AUC too). This is because the retraining with fairness-repaired samples, especially those whose labels are corrected to maximize predicted positives, introduces more recall-enhancing positive samples, which helps balance precision and recall, ultimately improving the overall quality. Thus, we conclude:

> ***RQ*$_4$*:*** CausalFT *generates test samples that make the model considerably more robust to fairness bugs compared with before (up to* 100% *cases) and the others (up to* 99% *cases). At the same time, the degradation on model quality is relatively small.*

5.5 Runtime Efficiency

5.5.1 Method. In **RQ**$_5$, we report the runtime for pre-testing analysis (e.g., causality analysis) and test sample generation, as the latter mainly depends on the base generator. We showcase the runtime required by the fastest and slowest generators for reaching 10,000 unique test samples. All statistics cover the results over all sensitive features and runs.

5.5.2 Result. From Figure 5a, we see that CausalFT is significantly more efficient than FairRF across all four models (DNN$_5$, DNN$_6$, LR, RF) in terms of pre-testing analysis, ranging from 8 to 78 seconds—up to 400+ seconds faster than FairRF. This is because when ranking the non-sensitive features, FairRF weights the Pearson correlations and these weights are, by default, learned by an SVM according to the learning rate. Training such an SVM is much more expensive than learning the causal graph in CausalFT.

For test sample generation (Figures 5b-5i), again, we note that CausalFT generally requires less time to generate 10,000 unique test samples than FairRF. This is because FairRF often selects the non-sensitive feature with many possible values, leading to a smaller space of perturbation compared with that of CausalFT. Thus, FairRF is more likely to generate redundant samples which cause longer runs. Yet, as we have shown, such a further reduced search space does not help FairRF to find more fairness bugs. The results also reveal that, compared with testing without CausalFT, CausalFT needs a slightly longer runtime to reach 10,000 unique test samples. This makes sense since CausalFT further reduces the search space in perturbation from that of testing without CausalFT, it is more likely to generate redundant samples, especially for datasets with a smaller search space. However, the extra runtime (including causal analysis) will vary depending on the type of model. It merely ranges from a few tens of seconds(under NeuronFair in LR) to a few hundred seconds(under SG in DNN$_6$). Overall, the worst case is ≈ 270 seconds longer under Law

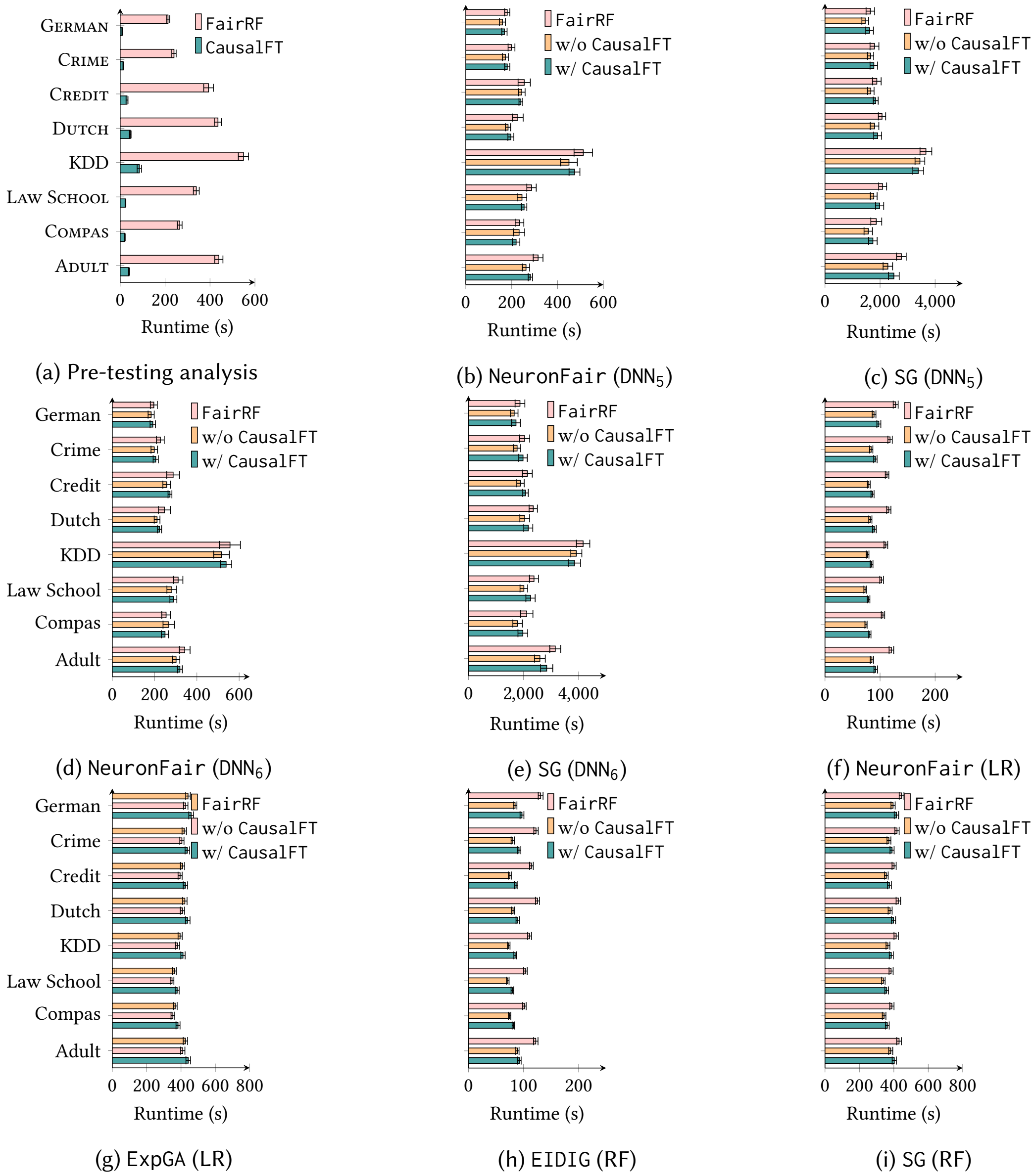


Fig. 5. Comparing runtime efficiency.NeuronFair (b) and SG (c) are the fastest and slowest generators for DNN$_5$; NeuronFair (d) and SG (e) are the fastest and slowest generators for DNN$_6$; NeuronFair (f) and ExpGA (g) are the fastest and slowest generators for LR; EIDIG (h) and SG (i) are the fastest and slowest generators for RF, respectively (mean/deviation over all possible sensitive features/runs).

School with SG for DNN$_6$ model. Importantly, as we have shown, CausalFT finds, e.g., hundreds more fairness bugs. Thus, we say:

> ***RQ_5:*** `CausalFT` *is much more efficient than* `FairRF` *but can cause slightly higher runtime overhead than testing without it. Yet, such extra overhead is acceptable for offline testing, especially considering the benefits it brings.*

6 Discussion

6.1 Modifying Base Generator

A strength of `CausalFT` is that it can be paired with different base generators, benefiting from their diverse perturbation strategies in fairness testing.

Specifically, `CausalFT` can benefit and integrate with base generators based on their designs:

- **Partial perturbation-based generators:** This includes, e.g., `ADF` and `EIDIG`, which naturally perturb all the non-sensitive features. Here, `CausalFT` can simply "trick" them to believe that both the concerned sensitive feature and the most causally relevant non-sensitive feature are part of the sensitive proportion, and hence to keep the other procedure unchanged. That is to say, `CausalFT` only seeds their local instances to form an expanded global candidate pool.
- **Full perturbation-based generators:** This includes, e.g., SG, such that all features are perturbed. In this case, we can simply feed the features other than the concerned sensitive feature and the most causally relevant non-sensitive feature into the generator for perturbation.

Integrating `CausalFT` with an arbitrary base generator requires minimal engineering effort. In our implementation, `CausalFT` operates as a wrapper around the generator's perturbation function, injecting causally guided perturbation strategies before calling the base generator. The same applies to partial perturbation-based generators such as `ADF` or full perturbation ones, e.g., `SG`. This requires approximately 50–100 lines of code changes depending on the generator's API, with integration time typically under 2 hours for a standard Python-based generator, including the time to comprehend its code structure. As such, `CausalFT` requires an acceptable amount of effort to be integrated into a new base generator.

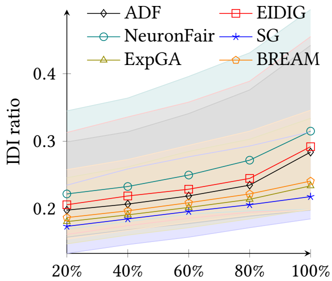
Indeed, some base generators require training data for their feature distribution learning [9, 31, 33, 63]. In those cases, they often leverage the training in their initialization stage while `CausalFT` benefits them at the perturbation stage. As such, using test data for fairness testing does not negatively affect the performance of these generators.

6.2 Sensitivity to k

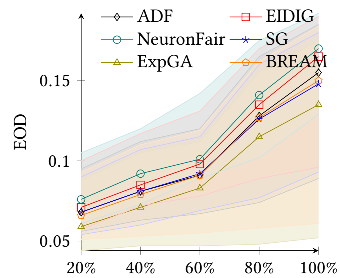
By default, we set $k = 100$, meaning that the causal graph in `CausalFT` is built by using 100% training data. From Figure 6, we see that although a smaller k would certainly lead to better efficiency, increasing k enables `CausalFT` to find more fairness bugs since a more reliable causal graph can be built, leading to more accurately estimated causal relationships/causal effects and hence better fairness testing.

One practical challenge is that some generators, such as ADF and FairRF, require access to the full training dataset to construct their internal models. As our focus is fairness testing on unseen data, we used test data to generate IDIs while ensuring no data leakage into model retraining or causal graph extraction. Empirically, we observed that using test data as the input for seed generation did not negatively affect base generator performance because their perturbation strategies operate independently of model training, and `CausalFT` focuses on input generation rather than model learning.

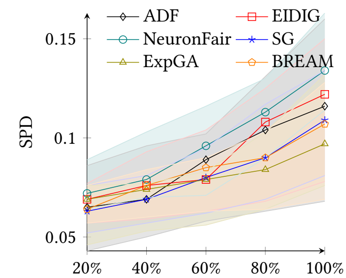
Key engineering tricks include aligning feature encoding schemas between `CausalFT` and base generators, and batching seed inputs to minimize invocation overhead when integrating with generators designed for large-batch sampling.



(a) IDI ratio



(b) EOD



(c) SPD

Fig. 6. Sensitivity of `CausalFT` to k over all models, datasets, possible sensitive features, generators, and runs.

6.3 Multiple Sensitive Features

While considering one concerned sensitive feature at a time is common [9, 28, 31, 39, 68, 74, 77, 79, 83], extending `CausalFT` with multiple sensitive features is also easy. There could be two ways for `CausalFT` to handle such a scenario with multiple sensitive features: the **direct** and **indirect** extension.

6.3.1 Indirect Extension. Since the presence of multiple sensitive features merely influences how discrimination instances are defined while `CausalFT` operates at the perturbation level of fairness testing, for the **indirect** way, one can perform the following:

(1) Pick a sensitive feature and run `CausalFT` as if that feature is the only sensitive one that is of concern.
(2) Repeat from 1) to cover all sensitive features.
(3) Combine all the samples generated and break the pairs; find the discrimination instances that fit the definition of multiple sensitive features.

6.3.2 Direct Extension. The **indirect** extension clearly requires no change on the internal mechanism of `CausalFT`, but relies on re-pairing the samples generated. Alternatively, we can also guide the perturbation from the beginning, i.e., the **direct** extension:

(1) Adjust the causal graph construction with multiple sensitive features as starting points.
(2) Modify the perturbation rules and the relaxed different fairness definition to accommodate combinations of sensitive features.

The above would enable `CausalFT` to uncover nuanced biases stemming from intersections of sensitive features, thereby providing a more comprehensive fairness testing therein.

It is clear that the key difference between the above two extensions in handling multiple sensitive features differs in terms of whether we conduct multiple runs, each targeting one sensitive feature individually and combine the results; or run once for all multiple sensitive features. Given sufficient budget, both ways should lead to similar outcomes. Yet, when the number of (unique) samples to be generated is small, the **direct** extension could be better in detecting hidden fairness bugs related to the interaction of all sensitive features.

6.4 Qualitative Analysis of `CausalFT`

Perturbation with the relaxed definition in `CausalFT` has enabled it to reveal many fairness bugs that are otherwise difficult to find. For example, when testing DNN under the Adult dataset, CausalFT finds the pair below which has never been identified by any other methods:

$$\mathbf{x'}_a = \{5, 0 \text{ (Male)}, 10, 3 \text{ (Separated)}, 4, 2, 1, 0, 0, 45, 0, 0, 0, 1\}$$

$$\mathbf{x'}_b = \{5, 1 \text{ (Female)}, 10, 4 \text{ (Married-civ-spouse)}, 4, 2, 1, 0, 0, 45, 0, 0, 0, 1\}$$

Clearly, it is an invalid pair under the true definition, since both the sensitive feature Gender (first index) and its most relevant non-sensitive feature Marital-status (third index) have different values. However, CausalFT then repairs by successfully finding another sample ($\mathbf{x''}_b$) from the testing dataset as:

$$\mathbf{x'}_b = \{5, 1 \text{ (Female)}, 10, 4, 4, 2, 1, 0, 0, 45, 0, 0, 0, 1\}$$

$$\mathbf{x''}_b = \{5, 0 \text{ (Male)}, 10, 4, 4, 2, 1, 0, 0, 45, 0, 0, 0, 1\}$$

In this corrected pair, only the Gender attribute differs, while all other features remain identical. Nevertheless, the model produces different predictions, indicating an individual discriminatory instance.

The above would be very difficult for the other approaches where only all the non-sensitive features (or all features) are to be perturbed, since the fairness bug can be more easily revealed if the value combination of both the sensitive feature and its most relevant non-sensitive counterpart is explicitly considered, which subsequently reduces the space of perturbation. Indeed, in practical scenarios, it is not uncommon that the gender of a person might strongly interact with his/her marital status to influence the decision made by an AI system.

6.5 Evidence-based Guidance for Generator Selection

Drawing on the findings from the experiments, we observe that although CausalFTcan significantly improve different base generators, the choice of which generator can indeed affect the results. In particular, NeuronFair is generally the best generator across the different models (all cases for DNN_5 and DNN_6; 16/18 cases for RF; 15/18 cases for LR). This is because NeuronFair leverages internal neuron activations and gradient signals to explore fine-grained decision boundaries and trace latent bias pathways, which are especially prevalent in complex and highly non-linear architectures such as deep neural networks. Even in statistical models such as RF and LR, these gradient-informed perturbations help navigate local decision regions more effectively and expose subtle fairness violations that heuristic search alone might miss.

However, there are some discrepancies between the models: NeuronFair is consistently the best across all datasets for DNN, while this might not always be the case for LR and RF. The key reason is that the generators are inherently influenced by the interplay between model and dataset features. Deep neural networks exhibit highly non-linear and fine-grained decision boundaries, which make them more sensitive to neuron-level perturbations and allow NeuronFair to exploit internal representations effectively. In contrast, models like RF and LR often construct monotonic decision surfaces, especially on datasets with lower feature dimensionality and strong feature correlations (e.g., German, Crime), where decision regions are easier to traverse. In such cases, the advantages of neuron-level guidance diminish, and gradient-agnostic strategies, e.g., EIDIG, may occasionally achieve comparable or even slightly better discriminatory instance discovery.

Drawing on the above evidence and findings, we outline a brief guideline for selecting the base generator while pairing with CausalFTas follows:

- In general, using NeuronFair is more reliable for DNN in all cases.
- NeuronFair is also generally the most robust choice for statistical machine learning models such as RF and LR when the characteristics of the target dataset are unknown.
- When it is possible to analyze the dataset, for models such as RF and LR, it is preferable to use EIDIG when the datasets have low feature dimensionality and strong feature correlations.

7 Threats to validity

Internal threats concern parameter settings. In this work, we follow the same settings from prior work [9, 77, 79, 81], including those for the generators, `FairRF`, and the fairness testing procedure. As for `CausalFT`, we examine its sensitivity to the most critical parameter k. Yet, unintended omission of information or options is always possible.

While we evaluated `CausalFT` on classical ML models and fully connected neural networks, the omission of other deep architectures (e.g., CNNs, RNNs) is a potential threat to external validity. However, it is worth noting that the modalities of such other DNNs, e.g., CNN and RNN can be highly different, i.e., the definition of image features could be different from tubular data. As such, applying the causality concept to those might require slight amendments to the definition, which might require future investigation. Indeed, fully connected neural networks remain the standard in prior fairness testing work [9, 31, 39, 50, 77, 79, 83].

Construct threats are related to the choice of metrics. We mitigate this by exploiting three individual/group fairness metrics of different types and covering diverse aspects. To ensure statistical significance, we use U-Test [65] and $\hat{\mathbf{A}}_{12}$ [64] to verify the results. Indeed, a more exhaustive study of a wider range of metrics can be part of our future work.

Finally, **external threats** to validity can come from the generalizability of the conclusions. To address this, we examine the eight most widely used datasets and six base generators with diverse characteristics. Those, together with the metrics considered, lead to 324 cases per model. Nevertheless, we agree that examining more diverse scenarios may prove fruitful.

8 Related Work

8.1 Causal Inference for Fairness

8.1.1 Causal Inference Approaches for Fairness Testing. There exists prior work that adopts causal inference fairness testing in AI systems [45, 55, 82]. Among others, Russell et al. [55] build a causal graph to estimate how sensitive features influence the label outcome. `FairQuant`[41] is an approach that aims to certify and quantify individual fairness of deep neural networks using symbolic interval analysis, focusing on altering different hyperparameter settings. Similarly, `Parfait-ML`[61] relies on search-based software to explore different hyperparameter configurations in machine learning libraries, identifying settings that improve fairness without significantly compromising accuracy. However, both of them search in the space of hyperparameters while `CausalFT` focuses on perturbation in the space of model input samples as the test cases.

`DICE` [50] is a tool that adopts an information-theoretic framework to quantify fairness defects by measuring the amount of protected information leakage using Shannon entropy and minimum entropy between the layers of a DNN model, which also follows a similar concept of causal inferences. Despite this, we see `DICE` as complementary to `CausalFT`: it can well serve as a base generator, benefiting from the advanced perturbation in `CausalFT` that preserves the interactions between sensitive features and the most causally relevant non-sensitive counterpart, similar to other base generators such as `ADF` and `EIDIG` which rely on the identical two-level search as `DICE`.

Overall, a key concept that distinguishes the above work and `CausalFT` is that they analyze the causal relationship between a (sensitive) feature and the label, or the in-between layers of a DNN model, while `CausalFT` extracts the causal information between sensitive and non-sensitive feature, capturing their interaction in jointly influencing the label, which is then used to guide perturbation.

8.1.2 Causal Inference Approaches for Fairness Measurement and Improvement. Recent studies have integrated causal inference into fairness analysis [38, 60, 75]. Among these, Wu et al. [66] propose a unified framework to quantify fairness via path-specific causal effects derived from causal graphs constructed using the PC algorithm. Its primary goal is to measure and explain unfairness by analyzing the causal pathways between sensitive attributes and outcomes, rather than guiding test generation via perturbations. `FairCFS` [47] is a causal feature selection method that constructs localized causal graphs to identify Markov blankets of class and sensitive features, thereby selecting features that block sensitive information transmission while preserving predictive performance. Zhang et al. [78] propose an adaptive fairness improvement framework leveraging causal graphs to inform retraining and data modification strategies for mitigating unfair outcomes. Their work focuses on fairness mitigation after model training though.

However, the above approaches directly focus on fairness measurement, explanation, or mitigation. In contrast, `CausalFT` is fundamentally different in its goal, where we aim for test case generation that reveals fairness bugs. This is achieved by integrating causal inference to identify non-sensitive features causally related to sensitive ones and systematically incorporating these into perturbation-based fairness testing.

8.2 Generator for Fairness Testing

A vast number of test sample generators have been proposed in recent years for fairness testing, mainly relying on different perturbation mechanisms. White-box generators are popular as they leverage internal information about AI model. `ADF` [79] is a generator that meticulously probes the model's sensitivity to input features with perturbation that leads to a biased outcome. Similarly, `EIDIG` [77] improves `ADF` with gradient information to guide the generation. `NeuronFair` [83] analyzes the biased neurons, from which it gains insight into the fundamental sources causing model bias. On the other hand, black-box generators can benefit from better exploration and diversity in testing generation. For example, `SG` [9] generates test samples based on the symmetry principle to guide the perturbation. `ExpGA` [31] combines model interpretation methods and genetic algorithms to search for biased regions. `BREAM` [39] builds a shadow model to simulate the behavior of the target AI model under testing and guides perturbation.

Galhotra et al. [33] detect individual discrimination by generating test cases differing only on sensitive features. While seminal for fairness testing, their approach does not utilize causal information to guide input generation, treating all non-sensitive features equally.

However, those generators either perturb the non-sensitive features or all features. `CausalFT`, in contrast, identifies the direct and most causally relevant non-sensitive feature to join with the concerned sensitive one, which is excluded in the perturbation to ensure their diversity and complex interactions are preserved. In particular, those generators are complementary to `CausalFT` such that they can serve as the underlying base generators.

8.3 Accounting Non-sensitive Features for Fairness Testing

Not much work has considered the usefulness of non-sensitive features in detecting fairness bugs. The most notable ones are `FairRF` [81] and `FairWS` [84], where the non-sensitive features are ranked based on correlation analysis and some fixed weights, which are then used to manipulate their contributions in AI model training to mitigate bias. Yet, they differ on how the ranking is exploited. Yan et al. [71] use clustering to divide datasets into groups, from which newly formed non-sensitive features can be used to retrain an AI model for better fairness.

Yet, the above ignores the causality in the datasets, particularly between non-sensitive and sensitive features, which, as we have shown, can be highly beneficial for fairness testing.

9 Conclusion

This paper proposes `CausalFT`, a generic, higher-level fairness testing framework using causal perturbation. This is achieved by extracting the most directly and causally relevant non-sensitive feature with respect to the concerned sensitive counterpart and injecting such information into the perturbation to guide the test sample generation. CausalFT is unique in the sense that it can be seamlessly paired with any perturbation-based generators. Extensive experiments on eight datasets, six generators, and three metrics demonstrate that CausalFT:

- significantly improves the ability of an existing generator to reveal fairness bugs;
- provides more reliable information on causality than simple correlation analysis;
- is capable of making AI system/model considerably more robust to bias with acceptable extra overhead.

In the future, we seek to combine the causal relationship between AI model hyperparameters and fairness with CausalFT, leading to richer causality-guided fairness testing while extending them for the cases of self-adaptive systems [19, 72].

Acknowledgments

This work was supported by an NSFC Grant (62372084) and a UKRI Grant (10054084).

References

[1] 1994. The german credit dataset. https://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29.
[2] 1998. The Law School dataset. https://www.kaggle.com/datasets/danofer/law-school-admissions-bar-passage.
[3] 2000. Census-Income (KDD). https://archive.ics.uci.edu/dataset/117/census+income+kdd.
[4] 2011. The Communities and Crime dataset. http://archive.ics.uci.edu/dataset/211/communities+and+crime+unnormalized.
[5] 2014. The Dutch Census of 2001 dataset. https://microdata.worldbank.org/index.php/catalog/2102.
[6] 2016. The compas dataset. https://github.com/propublica/compas-analysis. Retrieved on May 20, 2022.
[7] 2016. The default credit dataset. https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients.
[8] 2017. The adult census income dataset. https://archive.ics.uci.edu/ml/datasets/adult.
[9] Aniya Aggarwal, Pranay Lohia, Seema Nagar, Kuntal Dey, and Diptikalyan Saha. 2019. Black box fairness testing of machine learning models. In *Proceedings of the ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/SIGSOFT FSE 2019, Tallinn, Estonia, August 26-30, 2019*, Marlon Dumas, Dietmar Pfahl, Sven Apel, and Alessandra Russo (Eds.). ACM, 625–635. https://doi.org/10.1145/3338906.3338937
[10] Andrea Arcuri and Lionel C. Briand. 2011. Adaptive random testing: an illusion of effectiveness?. In *Proceedings of the 20th International Symposium on Software Testing and Analysis, ISSTA 2011, Toronto, ON, Canada, July 17-21, 2011*, Matthew B. Dwyer and Frank Tip (Eds.). ACM, 265–275. https://doi.org/10.1145/2001420.2001452
[11] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2021. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research* 50, 1 (2021), 3–44.
[12] Reuben Binns. 2018. Fairness in Machine Learning: Lessons from Political Philosophy. In *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 149–159. http://proceedings.mlr.press/v81/binns18a.html
[13] Reuben Binns. 2020. On the apparent conflict between individual and group fairness. In *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, Mireille Hildebrandt, Carlos Castillo, L. Elisa Celis, Salvatore Ruggieri, Linnet Taylor, and Gabriela Zanfir-Fortuna (Eds.). ACM, 514–524. https://doi.org/10.1145/3351095.3372864
[14] Sumon Biswas and Hridesh Rajan. 2020. Do the machine learning models on a crowd sourced platform exhibit bias? an empirical study on model fairness. In *ESEC/FSE '20: 28th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, Virtual Event, USA, November 8-13, 2020*, Prem Devanbu, Myra B. Cohen, and Thomas Zimmermann (Eds.). ACM, 642–653. https://doi.org/10.1145/3368089.3409704
[15] Jason Chan and Jing Wang. 2018. Hiring preferences in online labor markets: Evidence of a female hiring bias. *Management Science* 64, 7 (2018), 2973–2994.
[16] Pengzhou Chen and Tao Chen. 2026. PromiseTune: Unveiling Causally Promising and Explainable Configuration Tuning. In *48th IEEE/ACM International Conference on Software Engineering (ICSE)*. ACM.
[17] Pengzhou Chen, Tao Chen, and Miqing Li. 2024. MMO: Meta Multi-Objectivization for Software Configuration Tuning. *IEEE Trans. Software Eng.* 50, 6 (2024), 1478–1504. https://doi.org/10.1109/TSE.2024.3388910

[18] Pengzhou Chen, Jingzhi Gong, and Tao Chen. 2025. Accuracy can lie: On the impact of surrogate model in configuration tuning. *IEEE Transactions on Software Engineering* 51, 2 (2025), 548–580.
[19] Tao Chen. 2022. Lifelong Dynamic Optimization for Self-Adaptive Systems: Fact or Fiction?. In *IEEE International Conference on Software Analysis, Evolution and Reengineering, SANER 2022, Honolulu, HI, USA, March 15-18, 2022*. IEEE, 78–89. https://doi.org/10.1109/SANER53432.2022.00022
[20] Tao Chen and Miqing Li. 2021. Multi-objectivizing software configuration tuning. In *ESEC/FSE '21: 29th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, Athens, Greece, August 23-28, 2021*. ACM, 453–465. https://doi.org/10.1145/3468264.3468555
[21] Tao Chen and Miqing Li. 2023. The Weights Can Be Harmful: Pareto Search versus Weighted Search in Multi-objective Search-based Software Engineering. *ACM Trans. Softw. Eng. Methodol.* 32, 1 (2023), 5:1–5:40. https://doi.org/10.1145/3514233
[22] Tao Chen and Miqing Li. 2024. Adapting Multi-objectivized Software Configuration Tuning. *Proc. ACM Softw. Eng.* 1, FSE (2024), 539–561. https://doi.org/10.1145/3643751
[23] Zhenpeng Chen, Jie M. Zhang, Max Hort, Mark Harman, and Federica Sarro. 2024. Fairness Testing: A Comprehensive Survey and Analysis of Trends. *ACM Trans. Softw. Eng. Methodol.* 33, 5 (2024), 137:1–137:59. https://doi.org/10.1145/3652155
[24] David Maxwell Chickering. 2002. Optimal Structure Identification With Greedy Search. *J. Mach. Learn. Res.* 3 (2002), 507–554. https://jmlr.org/papers/v3/chickering02b.html
[25] Chengwen Du and Tao Chen. 2024. Contexts Matter: An Empirical Study on Contextual Influence in Fairness Testing for Deep Learning Systems. In *2024 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement, ESEM'24*. ACM.
[26] Mengnan Du, Fan Yang, Na Zou, and Xia Hu. 2021. Fairness in Deep Learning: A Computational Perspective. *IEEE Intell. Syst.* 36, 4 (2021), 25–34. https://doi.org/10.1109/MIS.2020.3000681
[27] Jannik Dunkelau and Manh Khoi Duong. 2022. Towards Equalised Odds as Fairness Metric in Academic Performance Prediction. *CoRR* abs/2209.14670 (2022). https://doi.org/10.48550/ARXIV.2209.14670 arXiv:2209.14670
[28] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. 2012. Fairness through awareness. In *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*, Shafi Goldwasser (Ed.). ACM, 214–226. https://doi.org/10.1145/2090236.2090255
[29] Ahmad-Reza Ehyaei, Golnoosh Farnadi, and Samira Samadi. 2023. Causal Fair Metric: Bridging Causality, Individual Fairness, and Adversarial Robustness. *CoRR* abs/2310.19391 (2023). https://doi.org/10.48550/ARXIV.2310.19391 arXiv:2310.19391
[30] Jan Eriksson and Visa Koivunen. 2004. Identifiability, separability, and uniqueness of linear ICA models. *IEEE Signal Process. Lett.* 11, 7 (2004), 601–604. https://doi.org/10.1109/LSP.2004.830118
[31] Ming Fan, Wenying Wei, Wuxia Jin, Zijiang Yang, and Ting Liu. 2022. Explanation-Guided Fairness Testing through Genetic Algorithm. In *44th IEEE/ACM 44th International Conference on Software Engineering, ICSE 2022, Pittsburgh, PA, USA, May 25-27, 2022*. ACM, 871–882. https://doi.org/10.1145/3510003.3510137
[32] Will Fleisher. 2021. What's Fair about Individual Fairness?. In *AIES '21: AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event, USA, May 19-21, 2021*, Marion Fourcade, Benjamin Kuipers, Seth Lazar, and Deirdre K. Mulligan (Eds.). ACM, 480–490. https://doi.org/10.1145/3461702.3462621
[33] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. 2017. Fairness testing: testing software for discrimination. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering, ESEC/FSE 2017, Paderborn, Germany, September 4-8, 2017*, Eric Bodden, Wilhelm Schäfer, Arie van Deursen, and Andrea Zisman (Eds.). ACM, 498–510. https://doi.org/10.1145/3106237.3106277
[34] Jingzhi Gong and Tao Chen. 2023. Predicting Software Performance with Divide-and-Learn. In *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/FSE 2023, San Francisco, CA, USA, December 3-9, 2023*. 858–870. https://doi.org/10.1145/3611643.3616334
[35] Jingzhi Gong and Tao Chen. 2024. Predicting Configuration Performance in Multiple Environments with Sequential Meta-Learning. *Proceedings of ACM Software Engineering* 1, FSE (2024), 359–382. https://doi.org/10.1145/3643743
[36] Jingzhi Gong, Tao Chen, and Rami Bahsoon. 2025. Dividable Configuration Performance Learning. *IEEE Trans. Software Eng.* 51, 1 (2025), 106–134. https://doi.org/10.1109/TSE.2024.3491945
[37] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (Eds.). 3315–3323. https://proceedings.neurips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html
[38] Zhenlan Ji, Pingchuan Ma, Shuai Wang, and Yanhui Li. 2023. Causality-Aided Trade-off Analysis for Machine Learning Fairness. *CoRR* abs/2305.13057 (2023). https://doi.org/10.48550/ARXIV.2305.13057 arXiv:2305.13057

[39] Weipeng Jiang, Chao Shen, Chenhao Lin, Jingyi Wang, Jun Sun, and Xuanqi Gao. 2023. Black-Box Fairness Testing with Shadow Models. In *Information and Communications Security - 25th International Conference, ICICS 2023, Tianjin, China, November 18-20, 2023, Proceedings (Lecture Notes in Computer Science, Vol. 14252)*, Ding Wang, Moti Yung, Zheli Liu, and Xiaofeng Chen (Eds.). Springer, 467–484. https://doi.org/10.1007/978-981-99-7356-9_28

[40] Zhimeng Jiang, Xiaotian Han, Chao Fan, Fan Yang, Ali Mostafavi, and Xia Hu. 2022. Generalized Demographic Parity for Group Fairness. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net. https://openreview.net/forum?id=YigKlMJwjye

[41] Brian Hyeongseok Kim, Jingbo Wang, and Chao Wang. 2024. FairQuant: Certifying and Quantifying Fairness of Deep Neural Networks. *CoRR* abs/2409.03220 (2024). https://doi.org/10.48550/ARXIV.2409.03220 arXiv:2409.03220

[42] Svetlana Kiritchenko and Saif M. Mohammad. 2018. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, *SEM@NAACL-HLT 2018, New Orleans, Louisiana, USA, June 5-6, 2018*, Malvina Nissim, Jonathan Berant, and Alessandro Lenci (Eds.). Association for Computational Linguistics, 43–53. https://doi.org/10.18653/V1/S18-2005

[43] Preethi Lahoti, Krishna P. Gummadi, and Gerhard Weikum. 2019. iFair: Learning Individually Fair Data Representations for Algorithmic Decision Making. In *35th IEEE International Conference on Data Engineering, ICDE 2019, Macao, China, April 8-11, 2019*. IEEE, 1334–1345. https://doi.org/10.1109/ICDE.2019.00121

[44] Miqing Li, Tao Chen, and Xin Yao. 2022. How to Evaluate Solutions in Pareto-Based Search-Based Software Engineering: A Critical Review and Methodological Guidance. *IEEE Trans. Software Eng.* 48, 5 (2022), 1771–1799. https://doi.org/10.1109/TSE.2020.3036108

[45] Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, and Yongfeng Zhang. 2021. Towards Personalized Fairness based on Causal Notion. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai (Eds.). ACM, 1054–1063. https://doi.org/10.1145/3404835.3462966

[46] Yanjun Li, Huan Huang, Xinwei Guo, and Yuyu Yuan. 2021. An Empirical Study on Group Fairness Metrics of Judicial Data. *IEEE Access* 9 (2021), 149043–149049. https://doi.org/10.1109/ACCESS.2021.3122443

[47] Zhaolong Ling, Enqi Xu, Peng Zhou, Liang Du, Kui Yu, and Xindong Wu. 2024. Fair Feature Selection: A Causal Perspective. *ACM Trans. Knowl. Discov. Data* 18, 7 (2024), 163. https://doi.org/10.1145/3643890

[48] Tasfia Mashiat, Xavier Gitiaux, Huzefa Rangwala, Patrick J. Fowler, and Sanmay Das. 2022. Trade-offs between Group Fairness Metrics in Societal Resource Allocation. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*. ACM, 1095–1105. https://doi.org/10.1145/3531146.3533171

[49] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2022. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* 54, 6 (2022), 115:1–115:35. https://doi.org/10.1145/3457607

[50] Verya Monjezi, Ashutosh Trivedi, Gang Tan, and Saeid Tizpaz-Niari. 2023. Information-Theoretic Testing and Debugging of Fairness Defects in Deep Neural Networks. In *45th IEEE/ACM International Conference on Software Engineering, ICSE 2023, Melbourne, Australia, May 14-20, 2023*. IEEE, 1571–1582. https://doi.org/10.1109/ICSE48619.2023.00136

[51] Carlos Mougan, Laura State, Antonio Ferrara, Salvatore Ruggieri, and Steffen Staab. 2023. Demographic Parity Inspector: Fairness Audits via the Explanation Space. *CoRR* abs/2303.08040 (2023). https://doi.org/10.48550/ARXIV.2303.08040 arXiv:2303.08040

[52] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453. https://doi.org/10.1126/science.aax2342

[53] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008*, Ying Li, Bing Liu, and Sunita Sarawagi (Eds.). ACM, 560–568. https://doi.org/10.1145/1401890.1401959

[54] Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. 2022. A survey on datasets for fairness-aware machine learning. *WIREs Data Mining Knowl. Discov.* 12, 3 (2022). https://doi.org/10.1002/WIDM.1452

[55] Chris Russell, Matt J. Kusner, Joshua R. Loftus, and Ricardo Silva. 2017. When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 6414–6423. https://proceedings.neurips.cc/paper/2017/hash/1271a7029c9df08643b631b02cf9e116-Abstract.html

[56] Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti J. Kerminen. 2006. A Linear Non-Gaussian Acyclic Model for Causal Discovery. *J. Mach. Learn. Res.* 7 (2006), 2003–2030. http://jmlr.org/papers/v7/shimizu06a.html

[57] Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O. Hoyer, and Kenneth Bollen. 2011. DirectLiNGAM: A Direct Method for Learning a Linear Non-Gaussian Structural Equation Model. *J. Mach. Learn. Res.* 12 (2011), 1225–1248. https://doi.org/10.5555/1953048.2021040

[58] Ricardo Silva. 2024. Counterfactual Fairness Is Not Demographic Parity, and Other Observations. *CoRR* abs/2402.02663 (2024). https://doi.org/10.48550/ARXIV.2402.02663 arXiv:2402.02663

[59] Edward A. Small, Kacper Sokol, Daniel Manning, Flora D. Salim, and Jeffrey Chan. 2023. Equalised Odds is not Equal Individual Odds: Post-processing for Group and Individual Fairness. *CoRR* abs/2304.09779 (2023). https://doi.org/10.48550/ARXIV.2304.09779 arXiv:2304.09779

[60] Cong Su, Guoxian Yu, Yongqing Zheng, Jun Wang, Zhengtian Wu, Xiangliang Zhang, and Carlotta Domeniconi. 2024. Causality-Based Fair Multiple Decision by Response Functions. *ACM Trans. Knowl. Discov. Data* 18, 3 (2024), 61:1–61:23. https://doi.org/10.1145/3632529

[61] Saeid Tizpaz-Niari, Ashish Kumar, Gang Tan, and Ashutosh Trivedi. 2022. Fairness-aware Configuration of Machine Learning Libraries. In *44th IEEE/ACM 44th International Conference on Software Engineering, ICSE 2022, Pittsburgh, PA, USA, May 25-27, 2022*. ACM, 909–920. https://doi.org/10.1145/3510003.3510202

[62] Florian Tramer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, and Huang Lin. 2017. Fairtest: Discovering unwarranted associations in data-driven applications. In *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 401–416.

[63] Sakshi Udeshi, Pryanshu Arora, and Sudipta Chattopadhyay. 2018. Automated directed fairness testing. In *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering, ASE 2018, Montpellier, France, September 3-7, 2018*, Marianne Huchard, Christian Kästner, and Gordon Fraser (Eds.). ACM, 98–108. https://doi.org/10.1145/3238147.3238165

[64] András Vargha and Harold D Delaney. 2000. A Critique and Improvement of the CL Common Language Effect Size Statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics* 25 (2000), 101 – 132. https://api.semanticscholar.org/CorpusID:120137017

[65] Frank Wilcoxon. 1992. Individual comparisons by ranking methods. In *Breakthroughs in statistics: Methodology and distribution*. Springer, 196–202.

[66] Yongkai Wu, Lu Zhang, Xintao Wu, and Hanghang Tong. 2019. PC-Fairness: A Unified Framework for Measuring Causality-based Fairness. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 3399–3409. https://proceedings.neurips.cc/paper/2019/hash/44a2e0804995faf8d2e3b084a1e2db1d-Abstract.html

[67] Zezhen Xiang, Jingzhi Gong, and Tao Chen. 2026. Dually Hierarchical Drift Adaptation for Online Configuration Performance Learning. In *48th IEEE/ACM International Conference on Software Engineering (ICSE)*. ACM.

[68] Yisong Xiao, Aishan Liu, Tianlin Li, and Xianglong Liu. 2023. Latent Imitator: Generating Natural Individual Discriminatory Instances for Black-Box Fairness Testing. In *Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis, ISSTA 2023, Seattle, WA, USA, July 17-21, 2023*, René Just and Gordon Fraser (Eds.). ACM, 829–841. https://doi.org/10.1145/3597926.3598099

[69] Gangda Xiong and Tao Chen. 2025. CoTune: Co-evolutionary Configuration Tuning. In *40th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE.

[70] Depeng Xu, Yongkai Wu, Shuhan Yuan, Lu Zhang, and Xintao Wu. 2019. Achieving Causal Fairness through Generative Adversarial Networks. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, Sarit Kraus (Ed.). ijcai.org, 1452–1458. https://doi.org/10.24963/IJCAI.2019/201

[71] Shen Yan, Hsien-Te Kao, and Emilio Ferrara. 2020. Fair Class Balancing: Enhancing Model Fairness without Observing Sensitive Attributes. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, Mathieu d'Aquin, Stefan Dietze, Claudia Hauff, Edward Curry, and Philippe Cudré-Mauroux (Eds.). ACM, 1715–1724. https://doi.org/10.1145/3340531.3411980

[72] Yulong Ye, Tao Chen, and Miqing Li. 2025. Distilled Lifelong Self-Adaptation for Configurable Systems. In *47th IEEE/ACM International Conference on Software Engineering, ICSE 2025, Ottawa, ON, Canada, April 26 - May 6, 2025*. IEEE, 1333–1345. https://doi.org/10.1109/ICSE55347.2025.00094

[73] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. 2017. Fairness Constraints: Mechanisms for Fair Classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA (Proceedings of Machine Learning Research, Vol. 54)*, Aarti Singh and Xiaojin (Jerry) Zhu (Eds.). PMLR, 962–970. http://proceedings.mlr.press/v54/zafar17a.html

[74] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating Unwanted Biases with Adversarial Learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2018, New Orleans, LA, USA, February 02-03, 2018*, Jason Furman, Gary E. Marchant, Huw Price, and Francesca Rossi (Eds.). ACM, 335–340. https://doi.org/10.1145/3278721.3278779

[75] Dacao Zhang, Kun Zhang, Le Wu, Mi Tian, Richang Hong, and Meng Wang. 2024. Path-Specific Causal Reasoning for Fairness-aware Cognitive Diagnosis. *CoRR* abs/2406.03064 (2024). https://doi.org/10.48550/ARXIV.2406.03064 arXiv:2406.03064

[76] Jie M. Zhang and Mark Harman. 2021. "Ignorance and Prejudice" in Software Fairness. In *43rd IEEE/ACM International Conference on Software Engineering, ICSE 2021, Madrid, Spain, 22-30 May 2021*. IEEE, 1436–1447. https://doi.org/10.1109/ICSE43902.2021.00129

[77] Lingfeng Zhang, Yueling Zhang, and Min Zhang. 2021. Efficient white-box fairness testing through gradient search. In *ISSTA '21: 30th ACM SIGSOFT International Symposium on Software Testing and Analysis, Virtual Event, Denmark, July 11-17, 2021*, Cristian Cadar and Xiangyu Zhang (Eds.). ACM, 103–114. https://doi.org/10.1145/3460319.3464820

[78] Mengdi Zhang and Jun Sun. 2022. Adaptive fairness improvement based on causality analysis. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/FSE 2022, Singapore, Singapore, November 14-18, 2022*, Abhik Roychoudhury, Cristian Cadar, and Miryung Kim (Eds.). ACM, 6–17. https://doi.org/10.1145/3540250.3549103

[79] Peixin Zhang, Jingyi Wang, Jun Sun, Guoliang Dong, Xinyu Wang, Xingen Wang, Jin Song Dong, and Ting Dai. 2020. White-box fairness testing through adversarial sampling. In *ICSE '20: 42nd International Conference on Software Engineering, Seoul, South Korea, 27 June - 19 July, 2020*, Gregg Rothermel and Doo-Hwan Bae (Eds.). ACM, 949–960. https://doi.org/10.1145/3377811.3380331

[80] Tianxiang Zhao, Enyan Dai, Kai Shu, and Suhang Wang. 2021. You Can Still Achieve Fairness Without Sensitive Attributes: Exploring Biases in Non-Sensitive Features. *CoRR* abs/2104.14537 (2021). arXiv:2104.14537 https://arxiv.org/abs/2104.14537

[81] Tianxiang Zhao, Enyan Dai, Kai Shu, and Suhang Wang. 2022. Towards Fair Classifiers Without Sensitive Attributes: Exploring Biases in Related Features. In *WSDM '22: The Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event / Tempe, AZ, USA, February 21 - 25, 2022*, K. Selcuk Candan, Huan Liu, Leman Akoglu, Xin Luna Dong, and Jiliang Tang (Eds.). ACM, 1433–1442. https://doi.org/10.1145/3488560.3498493

[82] Xuan Zhao, Klaus Broelemann, Salvatore Ruggieri, and Gjergji Kasneci. 2023. Causal Fairness-Guided Dataset Reweighting using Neural Networks. In *IEEE International Conference on Big Data, BigData 2023, Sorrento, Italy, December 15-18, 2023*, Jingrui He, Themis Palpanas, Xiaohua Hu, Alfredo Cuzzocrea, Dejing Dou, Dominik Slezak, Wei Wang, Aleksandra Gruca, Jerry Chun-Wei Lin, and Rakesh Agrawal (Eds.). IEEE, 1386–1394. https://doi.org/10.1109/BIGDATA59044.2023.10386836

[83] Haibin Zheng, Zhiqing Chen, Tianyu Du, Xuhong Zhang, Yao Cheng, Shouling Ji, Jingyi Wang, Yue Yu, and Jinyin Chen. 2022. NeuronFair: Interpretable White-Box Fairness Testing through Biased Neuron Identification. In *44th IEEE/ACM 44th International Conference on Software Engineering, ICSE 2022, Pittsburgh, PA, USA, May 25-27, 2022*. ACM, 1519–1531. https://doi.org/10.1145/3510003.3510123

[84] Huaisheng Zhu, Enyan Dai, Hui Liu, and Suhang Wang. 2023. Learning fair models without sensitive attributes: A generative approach. *Neurocomputing* 561 (2023), 126841. https://doi.org/10.1016/J.NEUCOM.2023.126841