**Introduction**

Virtually every industry that uses predictive policies or generates opportunities is susceptible to unfairness, even in unexpected cases like a product promotion or discount given to only select customers. Machine learning (ML) can have legal and ethical consequences if their outputs lead to demographic decision imbalances. Industries use pre-trained ML models to feed input data to produce an output – a decision for business actions. These models are composed of learned parameters that affect the probability of a decision. Decision bias can have multiple causes, such as if a model is trained on data containing minorities, ‘representation disparity’ can occur for minority groups. [[5]](#five) Testing the fairness of these parameters in model fairness testing (MFT) is typically done *without* inspecting the parameters themselves. This makes the models a ‘black box,’ meaning that testing them entails giving them input and observing the output, without knowing how or why the model came to its output. Li et al. states model fairness emphasizes that “similar individuals should be treated similarly” [[1]](#one).

**Related Work**

While group and individual notions of fairness will be discussed in a contradictory manner, it should be noted that this is in regard to model fairness detection, not in a broader phenomenological sense. [[8]](#eight)

Past research tests fairness by perturbating over every sensitive group for each known sample to create an ‘ambiguity set’ as an output probability distribution. This group approach is known as Distributionally Robust Optimization (DRO). DRO has mainly been used to guard against models overfitting to majority groups by focusing on the worst-performing groups, ensuring no group is left behind. [[5]](#five) However, while DRO addresses group-level performance, this broad-stroked methodology misses unfairness on the individual level by aggregating outcomes rather than checking fairness on each individual.

Hardt et al. originally proposed ‘group fairness’ to request equalized odds from algorithms across sensitive [[4]](#four). However, equalized odds only enforce equal false positive and false negative rates across groups, without considering whether similar individuals from different groups receive the same decisions. [[7]](#seven)

In another critique of group fairness definitions, Gölz et al. proposed that a set of axioms from axiomatic fairness theory (resource monotonicity, population monotonicity, consistency) must be required to avoid a paradox of equalized odds. They discovered a paradox between simultaneously maintaining fairness and the population monotonicity axiom. A real-world situation could arise where “we need to decrease a student’s financial aid because another student declined to accept aid.” They admit that “it is hard to justify and explain the design of allocation algorithms that behave in such counter-intuitive ways.” [[9]](#nine) *Even if equalized odds is mathematically satisfied (fair) on principle fairness axioms, seemingly unfair situations reveal a deeper structural issue.*

Therefore, this report pivots from group definitions toward the individual level, which provides a more robust path to revealing bias. Wachter et al. introduced the notion of counterfactual explanations that find the minimal altered instance that flips the model’s decision [[2]](#two). This notion was formally defined by Kusnar et al. as a “counterfactual world” where an individual differs only in their sensitive, ‘protected’ attribute *S* [[3]](#three). These counterfactual approaches reveal nuanced individual instances of discrimination. Common counterfactual methodologies include latent variable models, gradient-based search, or optimization based techniques.

This paper will show how an empirically grounded local neighbour (LN) approach was implemented to reveal instances of AI model discrimination through individual counterfactual explanations. This added local approach allows for a fast practical blend of both group-level and individual-level consideration to explore the decision boundary of protected features using synthetic but empirically-preserving perturbations.

**Solution**

Given a datapoint with a sensitive feature *s*, the DRO group paradigm changes *S* while keeping the non-sensitive features fixed. This can jump the datapoint out of the real distribution for that group because correlations that would be typical of the group are not the same as with s. Consequently, may lie far outside the true data distribution, especially if s is strongly correlated with other features. If the model would output a different target value after , this would be an unfair false positive. It is important to maintain correlations in the data because AI model fairness testing does not aim to exclude real world differences in demographics. Rather, fairness is tested by reducing the variability along every axis except the sensitive features.

The proposed approach, demonstrated in Figure [1](#figureOne), is neighbour perturbation of data points to generate pairs of synthetic individuals to seek counterfactual expressions, as opposed to a time-expensive similarity-based search. It is entirely plausible to use only real pairs in the dataset, but finding similar pairs can be computationally expensive with larger datasets with higher dimensionality. Additionally, since the model’s generalization capacity is unknown, it is equally meaningful to test the model’s fairness on unseen data. In this methodology, unseen data is synthetically devised by pairing individuals and perturbing their non-sensitive features with consistent direction (direction of similarity) and intent.

By shifting non-sensitive features along the axis of known values, the result is a pair of more similar individuals who only differ in their sensitive features. This methodology avoids assuming causal relationships between features and relies on proximity. As a result, the distribution of the data is better maintained than with flipping, as shown in Figure [2](#figureTwo). However, the problem with neighbour perturbation, shown in Table Z, is that it is not guaranteed that instances will be similar depending on how far apart they are.

A diagram of values and values

AI-generated content may be incorrect.

Figure 1: Example perturbation on features from processed\_law\_school\_cleaned.csv

Before testing fairness, all unique values for each non-sensitive feature are stored and sorted prior to fairness testing. Suppose and belong to two different sensitive groups *.* Then a small perturbation is made on non-sensitive features:

where it is guaranteed, each feature is replaced by a neighboring value. Each new feature remains close to the distribution of its respective group, within the known distribution range.

A graph with a purple line

AI-generated content may be incorrect.

Figure 2: A simple 2-dimensional fake example showing that local perturbation yields more empirically similar synthetic data.

Setup

This section needs to include the experimental setup and procedure, such as some brief

discussion of systems/projects, parameter setting, evaluation metrics, what baseline to compare,

and statistical analysis used.

* The experimental setup is contained by the parameter dictionary in run.py.
* The model is compared to baseline random search in Table [1](#tableOne).
* The IDI Ratio is the ratio of individual discriminatory instances to the total number of instances. tested, and the total number of tested instances is expressed as a percent of the dataset size.

Experiments

* The solution could be assessed on any dataset with sensitive features, in a real-world scenario where discrimination can arise.
* The quality of the solution (revealed IDI) is evaluated based on the Spearman similarity score. It is important to note that only real instances were drawn from the dataset, then shifted along their unique feature values axes, as opposed to generating completely new instances. Figure [4](#figureFour) shows how the distribution is maintained across the entire dataset for an individual feature, whereas the Spearman score indicates the maintained relationship of the distributions after neighbour perturbation.
* While the solution did not reveal any IDI instances on 2 datasets, the baseline was significantly outperformed on all others.

Table 1: Experiments were run with a combination\_length = 2, max\_combo\_pairs = 20 and pairs\_per\_combo\_pair\_pctg = 10.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Dataset Name** | **% total data** | ***S* features** | **Target** | **IDI Ratio (LN)** | **IDI Ratio (RS)** |
| Compas | 400.33% | Sex, Age, Race | COMPAS Score | 38.28% | 11.7% |
| German | 400% | Sex, Age | Credit Rating | 28.1% | 1.3% |
| Adult | 400.07% | Sex, Age, Race | Class Label | 0.0% | 0.0% |
| Credit | 300% | Sex, Age | Class | 33.6% | 8.2% |
| Dutch | 400% | Sex, Age | Economic status | 0% | 0.0% |
| KDD | 400% | Sex, Age, Race | Income | 27.1% | 0.6% |
| Law School | 120% | Sex, Race | LSAT | 46.13% | 3.6% |

Table 2: Law School Datasets, differing only in race (Max=1)

|  |  |  |  |
| --- | --- | --- | --- |
|  | # race-differing pairs | Mean Spearman Correlation | Min |
| Real | 2793 | 0.826 | 0.279 |
| Synthetic | 8320 | 0.751 | -0.111 |

A graph with different colored lines

AI-generated content may be incorrect.

Figure 4: The distributions are tightly maintained for perturbated features of the German dataset using 400% of dataset size.

Reflection  
  
A local neighbour perturbation can create unrealistic samples, despite feature correlations remaining intact. An unrealistic perturbation could arise in a situation where a feature has very few unique values, so its synthetic counterpart could be considered unexpected by the model. An IDI occurring for unnatural synthetic data doesn’t suggest the model is bias or overfitted to unseen data, but can occur because a more complex, high-dimensional pattern was broken that the Spearman test did not reflect in Table [2](#tableTwo).

A Generative Adversarial Network could be tasked with generating even more natural synthetic pairings. Browsing for this idea showed one such study deploys a framework called **L**atent **Imi**tator (LIMI). With the help of a GAN, they learn the latent decision boundary of the model whose parameters are unknown, and therein sample latent instances in this region to uncover more discriminatory instances (*X9.42 instances*) with more naturalness (*+19.65%*) compared to baselines. [[6]](#six)

Conclusion

This report has comprehensively researched the case of group and individual notions of statistical fairness testing. Early on, it became apparent that individual counterfactual explanations can become extremely expensive to find, and generating synthetic data can give misleading results. The Local Neighbour solution found significantly outperformed the baseline random search on all but two instances. This solution loses some of the correlations between features, but the Spearman test showed that it was adequately maintained relative to the magnitude of IDI ratios uncovered.

Artifact  
  
The code and raw results can be found [here](https://github.com/KSkert/ISE_Fairness_Test). (https://github.com/KSkert/ISE\_Fairness\_Test)

References

[1] Li, P., Xia, E., & Liu, H. (2023, July). Learning antidote data to individual unfairness. In *International Conference on Machine Learning* (pp. 20168-20181). PMLR.  
  
[2] Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, *31*, 841.  
  
[3] Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. *Advances in neural information processing systems*, *30*.

[4] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, *29*.

[5] Hashimoto, T., Srivastava, M., Namkoong, H., & Liang, P. (2018, July). Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning* (pp. 1929-1938). PMLR.

[6] Xiao, Y., Liu, A., Li, T., & Liu, X. (2023, July). 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* (pp. 829-841).

[7] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012, January). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference* (pp. 214-226).

[8] Binns, R. (2020, January). On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 514-524).

[9] Gölz, P., Kahng, A., & Procaccia, A. D. (2019). Paradoxes in fair machine learning. *Advances in Neural Information Processing Systems*, *32*.