Synthetic Dataset Generation for Fairer Unfairness Research

Lan Jiang
lanj3@illinois.edu
School of Information Sciences,
University of Illinois
Urbana-Champaign
Champaign, Illinois, USA

Clara Belitz
cbelitz2@illinois.edu
School of Information Sciences,
University of Illinois
Urbana-Champaign
Champaign, Illinois, USA

Nigel Bosch
pnb@illinois.edu
School of Information Sciences and
Department of Educational
Psychology, University of Illinois
Urbana-Champaign
Champaign, Illinois, USA

ABSTRACT

Recent research has made strides toward fair machine learning. Relatively few datasets, however, are commonly examined to evaluate these fairness-aware algorithms, and even fewer in education domains, which can lead to a narrow focus on particular types of fairness issues. In this paper, we describe a novel dataset modification method that utilizes a genetic algorithm to induce many types of unfairness into datasets. Additionally, our method can generate an unfairness benchmark dataset from scratch (thus avoiding data collection in situations that might exploit marginalized populations), or modify an existing dataset used as a reference point. Our method can increase the unfairness by 156.3% on average across datasets and unfairness definitions while preserving AUC scores for models trained on the original dataset (just 0.3% change, on average). We investigate the generalization of our method across educational datasets with different characteristics and evaluate three common unfairness mitigation algorithms. The results show that our method can generate datasets with different types of unfairness, large and small datasets, different types of features, and which affect models trained with different classifiers. Datasets generated with this method can be used for benchmarking and testing for future research on the measurement and mitigation of algorithmic unfairness.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; *Genetic algorithms.*

KEYWORDS

datasets, student data, fair machine learning, data generation

ACM Reference Format:
Lan Jiang, Clara Belitz, and Nigel Bosch. 2024. Synthetic Dataset Generation for Fairer Unfairness Research. In *The 14th Learning Analytics and Knowledge Conference (LAK '24), March 18–22, 2024, Kyoto, Japan.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3636555.3636868

LAK '24, March 18–22, 2024, Kyoto, Japan

ACM ISBN 979-8-4007-1618-8/24/03…$15.00
https://doi.org/10.1145/3636555.3636868

1 INTRODUCTION

Machine learning is increasingly used in education, such as when predicting dropouts to facilitate advising processes [5] or for personalized learning to foster improved academic outcomes [67]. The development of predictive models for these systems often relies on curated datasets. These systems learn to identify extant patterns in those data that provide valuable insights for fast, scalable decision making [64]. In the process of curating the datasets, however, some students may be underrepresented, or be accurately represented as only a small fraction of the data. Moreover, the data itself can represent unfair trends in education or even exacerbate them [36]. Thus, even though machine learning presents novel solutions and tools for both public and private use, these models are not always implemented in equitable ways. For example, African American men tend to receive a lower score from an automatic essay scoring system than a human rater[17]. As such, concerns about the potential impacts of these applications have been repeatedly documented [6, 10, 53, 55, 65]. While a variety of algorithmic and sociotechnical approaches to fairness have been proposed, a relatively small number of datasets are commonly examined to evaluate these concerns [14, 61]. In the field of education, even fewer datasets are available for examining fairness. Furthermore, the construction of datasets presents several issues, such as including features like student sex and job of students' parents in the UCI student performance dataset, and incorporating caste (an old system where people are born to different groups that determine their social status) as a feature in the publically available, UCI student academic dataset [25]. In this paper, we describe a novel dataset modification and generation method that utilizes a genetic algorithm to induce many types of unfairness into datasets, allowing a broader range of research on unfairness mitigation.

Existing works have demonstrated that even state-of-the-art models are prone to making biased predictions with respect to a variety of dimensions of identity [6, 26, 33, 65]. Generally, bias in machine learning is defined in relation to existing demographic and/or personal identifiers and is measured through observing disparate outcomes in some element of the prediction process, such as the accuracy or desirability of predictions across relevant subgroups [55]. Categories used for comparison are commonly demographic, including sex and race, but other identities, such as socioeconomic status or age, can be similarly examined [11, 55, 70, 73].

In machine learning, data largely impact the final outcomes and even shape the questions we are able to ask [9]. As such, there has been a growing discussion about where these data come from and how well they represent the problems being addressed [60]. Concerns include issues of privacy, a lack of generalizability, and limited

documentation [7, 34, 50, 60]. Similarly, with the rapid growth in educational AI, many applications are driven by learning-related data. Though these systems aimed to enhance the quality of education students received, they were still at risk of amplifying the unfairness reflected in societies. Additionally, the use of demographic data in biased training datasets has been specifically critiqued [5]. As a result, fair educational AI focuses on the debiasing of unfairness. Recent work has surfaced limitations with the benchmark datasets used to provide standardized fairness comparisons across algorithms [28]. Thus, there is a need for a standardized way of generating test data for classification problems, particularly in the field of algorithmic fairness.

In this work, we propose using genetic algorithms to generate unfair benchmark datasets with particular unfairness properties, for the purpose of researching bias measurement and mitigation strategies. We explored the usefulness and generalization of our proposed method by addressing the following research questions:

- RQ1: Can genetic algorithms be modified to create a method for generating synthetic educational datasets that capture unfairness while maintaining accuracy?
- RQ2: For what kind of classification problems do data generated by genetic algorithms accurately capture unfairness? Does the difficulty of the problem affect the performance of this generation method?
- RQ3: How do existing fair algorithmic approaches perform given generated unfairness benchmarks, and for which fairness definitions?

We show that data can be generated to induce unfairness with respect to specific unfairness metrics, creating a replicable way of testing whether unfairness mitigation methods account for, and subsequently remove, specific patterns of bias. These generated datasets can mimic specific real-world situations of unfairness while avoiding some of the pitfalls of existing fairness benchmarks. The code used to generate data is available at https://github.com/lanj/unfair_dataset_generation.

2 RELATED WORK

Data are an essential part of machine learning, providing domain specificity and shaping what problem(s) an algorithm can learn to solve [9]. While domain generalizability and transfer learning are research areas of interest, it still represents a difficult problem [80], demonstrating the need for training data which represent the problem being studied. The role of data extends to how we measure fairness in machine learning systems. Fairness does not exist in a vacuum, since it is measured in relation to labels and individual student, teacher, or group categories present in the data. Datasets may pose challenges to accurately assessing fairness in that they can lack predictive information of interest, be sparse, or neglect relevant demographic categories, among other issues [12, 46, 66].

2.1 Shortcomings in existing benchmark datasets

To compare results across varying approaches, a small subset of datasets are commonly examined as fairness benchmarks across many machine learning projects [9, 34, 50]. Related work has surfaced a variety of issues concerning these common benchmark datasets used for machine learning, however, ranging from privacy concerns to problems of domain specificity to the way categories are defined [5, 7, 19, 28, 34, 66, 73]. One of the most well-known benchmarks, COMPAS, concerns recidivism predictions for people incarcerated in Florida. As pointed out by Bao et al. [7], the use of this criminal justice data for fairness benchmarking ignores the sociotechnical, contextual grounding relevant to this type of risk prediction. In addition, the COMPAS dataset may not even capture what we think it does, since it focuses on "re-arrest" rather than "re-offense" [7]. Researchers have similarly noted issues with the *Adult* dataset and called for the end of its usage [28]. Additionally, researchers in the educational AI field are concerned about designing and collecting data: whether the dataset is demographically representative of the societies, and whether demographics cause differences in learning variables [6, 24]. In general, there is a growing critique of the focus on a small number of benchmarks for fairness and quality (e.g. representative of different contexts, documentation, and design) of the datasets[7, 10, 34].

Privacy is an additional concern for data about people. Even when a dataset is "anonymized," it may be possible to identify individuals unless the data have been subjected to rigorous standards using algorithms that support differential privacy [15, 31]. Anonymizing data might also necessitate removing the fine-grained information required for exploring fairness with respect to small groups. This may generally disadvantage identities that are underrepresented in the data, as well as potentially preclude the ability to consider intersectionality [73].

2.2 Dataset generation in machine learning

While data sharing is important for transparency and reproducibility, a number of studies have identified how even anonymized data can disclose private information (e.g., membership inference [41, 63, 68] and model inversion [37, 57]). One type of solution is the use of synthetic datasets, which mimic statistical properties of the original data without including identifying information [1, 30]. Synthetic data can be shared and reproduced without risking the privacy of students or other individuals represented in the data, and have the additional benefit of testing different hypotheses by representing specific patterns or scenarios at low cost. This is especially useful for edge cases that may not be easy to capture [47, 56], allowing researchers to conduct experiments even when empirical data is largely unavailable.

Various algorithms have been proposed to generate synthetic data. These algorithms have a shared goal of trying to model relations among variables that were present in the original dataset [3]. Most algorithms calculate distributions of variables of interest and develop probabilistic models (e.g., mixture of Gaussians, multinomial feature distributions) that generate data [2, 23, 52, 69]. Machine learning techniques like Bayesian networks [79], support vector machines [29], and random forests [20] have been used to generate synthetic data. More recently, synthetic data generation has been boosted by deep learning models, especially generative adversarial networks (GANs) [39], which were originally proposed to simulate realistic synthetic images but have since been used for a variety of domains [22, 32, 48, 77], including fairness-related research [54].

For example, Xu et al. [76] and van Breugel et al. [71] have introduced methods to use GANs to generate tabular, fair synthetic data in order to train fair models, a novel approach to preprocessing. Generating synthetic data for fairness research is a growing topic, with recent work using structural equation modeling to generate biased datasets [8], using synthetic data to examine root causes of bias in datasets [21], and demonstrating that synthetic data can unintentionally create bias [40].

However, these algorithms target a specific bias measurement and require the data to conform to a specific distribution. In the realm of education, data usually consists of different types of variables, including categorical, binary, and continuous. Biases in those data are driven by complex reasons, making education-specific metrics and resources required. Though a single dataset is unlikely to capture the entirety of those reasons, ranging from parental support to teacher enthusiasm to student engagement, having strong training data can allow our predictive models to provide as accurate of insights as possible. In this study, our aim is to modify genetic algorithms to handle diverse datasets and adapt various unfairness definitions, using a transferable methodology for generating novel training data.

3 METHOD

3.1 Genetic algorithm

Genetic algorithms are gradient-free optimization algorithms inspired by the process of natural selection [74]. Natural selection is the process of differential survival of individuals based on their traits, which is the mechanism of evolution [27]. Genetic algorithms take inspiration from this idea to generate fit "individuals" (i.e., optimized solutions) by iteratively going through four subprocesses: evaluation, selection, crossover, and mutation [74]. To begin, a genetic algorithm creates a set of random individuals (i.e., solutions). During each iteration, only a proportion of the fittest individuals from the population are selected to create the next set of offspring. These individuals are chosen according to the fitness function, which may have any form (including non-differentiable functions) as long as it can rank solutions. Next, pairs of the selected individuals are combined at randomly chosen crossover points to form new individuals. To mimic genetic drift over time, mutations are then introduced into the offspring. This process repeats until a termination condition is met, such as a certain number of iterations or a desired fitness. The fittest individual at the point of termination is then returned as the best solution found.

To adapt to the goal of generating unfairness benchmarks for learning analytics purposes, we modified the classical genetic algorithm as follows:

Initialization: In this phase, a large number of datasets are generated based on a reference dataset, rather than purely random data. For our experiments, we worked with both real-world and simulated reference datasets (Section 4). The reference dataset is used to generate a variety of datasets (i.e., individuals in a classical genetic algorithm). To create each individual dataset in the initial population, we sampled from the original distribution for each feature.

Evaluation: The goal of our method is not only (i) to have a high unfairness score, but also (ii) to retain important predictive patterns embedded in the original dataset. Thus, the relationships among features should be consistent in order to ensure the generated dataset's usefulness. We achieved this two-part goal by considering two terms in our fitness function: an unfairness measurement and a similarity measurement ($F = S_{unfairness} + S_{similarity}$). Which unfairness definitions are most applicable depends on the specific goal being pursued [13]; thus, different unfairness calculations are used in our experiments, and can be user-specified. We ensure a dataset's usefulness by maximizing the percentage of values that are the same between the generated dataset and the original reference dataset. Maintaining a high similarity between the generated and original data helps retain the important patterns in the dataset (e.g., predictive relationships between student behaviors and learning outcomes) while changing only what needs to differ for a particular type of bias to emerge.

- Similarity measurements: Similarity between two datasets can be measured in a variety of ways. Three common metrics are mean absolute error, mean squared error, and the percentage of matching values [78]. In our study, we calculated the percentage of matching values between the original reference and the generated datasets. Percentage was chosen because the range is consistent regardless of the magnitude or data types in the chosen datasets. Percentage similarity also makes no distributional assumptions about the datasets. Hence, we did not need to adjust the similarity measure for different datasets.
- Unfairness measurements: We tested three popular unfairness measurements [13, 62]: overall accuracy equality, statistical parity, and calibration equality. We choose the first two measurements because they represent substantially different types of fairness that cannot be achieved simultaneously. The third one offers a different perspective by measuring the unfairness of prediction probabilities, rather than categorical decisions. The mathematical calculations of them are presented in Section 3.2.

Selection: During selection, individuals from the current iteration are selected to be "parents" of the next iteration. We used roulette wheel selection [51], where all individuals are given a probability of being selected according to their fitness score. Conceptually, roulette wheel selection is represented by individuals with a higher fitness score having a larger region assigned to them on a circular wheel. Parents are selected by randomly choosing a fixed point on this wheel. Therefore, a fitter individual (i.e., better dataset) has a greater chance of being selected as a parent.

Crossover: In this phase, offspring are generated by combining genetic information (i.e., values in the dataset) from two parents chosen in the selection operation. For each pair of parent datasets, a point on the feature list is randomly selected. The features of the two parents are swapped at this crossover point to produce new datasets, the "offspring."

Mutation: Mutation is an operation that can maintain genetic diversity and help solutions escape local optima. After offspring have been produced, 0.2% of each offspring's values will be replaced. Mutation rates larger than this tend to lead to an essentially random search [58]. For our method, we first randomly pick which values will be replaced, then replace each of them with a value sampled

from the original distribution of each feature. The population for the next iteration is made up of these mutated datasets.

3.2 Unfairness Metrics

We enumerate the unfairness metrics used in our experiments in both mathematical and descriptive terms. For all three metrics described below, a higher score indicates a higher level of unfairness. All of the metrics range from 0 to 1. Note that TP represents the number of true positives, TN represents the number of true negatives, FP represents the number of false positives, and FN represents the number of false negatives. "True" here refers to the predictions being the same as the ground truth labels, while "false" refers to the predictions not being the same as the ground truth labels. The unfairness definitions used in this paper are as follows:

(1) *Calibration equality* is achieved when people with any predicted probabilities scores have an equal probability of being in a positive class. In our calculation, we calculated calibration scores for each group of people. The calibration scores are calculated by the difference between true and predicted probabilities of a certain number of instances, sorted by predicted probability. We set the number of instances in one bin as 20% of the number of instances in the whole dataset. The unfairness under this definition is calculated by the largest difference in calibration scores over the groups.
(2) *Overall accuracy equality* is achieved when the accuracy is equal for different groups of people. That is, $\frac{TP+TN}{TP+FP+TN+FN}$ is the same for all groups. In this case, true positives and true negatives are equally important. In cases with more than two possible groups, we calculated the unfairness score as the largest difference in accuracy across all groups (i.e., the highest-accuracy group compared to the lowest).
(3) *Statistical parity* is achieved when the distribution of predicted classes is the same across groups. That is, $\frac{TP+FP}{TP+FP+TN+FN}$ or $\frac{TN+FN}{TP+FP+TN+FN}$ is the same for all groups. The unfairness under this definition is calculated by the largest difference in the proportion of predicted positive instances across all groups.

4 DATASETS

In this section, we introduce the datasets used as reference datasets.

4.1 Simulated datasets

When starting from scratch, without a real-world dataset, we used the `make_classification` function in *Scikit-learn* [18] to generate a synthetic dataset. This function can generate a classification problem by setting the number of features, the number of informative (i.e., useful) features, the number of samples, and the difficulty of the problem. We then transformed the starting dataset to create two reference datasets:

(1) **Simulated dataset (1 feature type: continuous)** dataset contains 1,000 rows in total and 16 features, 10 of which are informative (i.e., correlated with labels and orthogonal to other features). The labels are binary: 0 and 1. We converted one additional feature to a binary format and then treated it as a *sensitive feature* – i.e., one which describes group membership.
(2) **Simulated dataset (3 feature types: continuous, binary, categorical)** dataset also contains 1,000 rows and 16 features, 10 of which are informative. To mimic real-world education datasets that include a mix of variable types—e.g., the law school dataset [75], which contains three categorical variables, three binary variables, and six continuous variables—we converted six features into binary features, one of which served as a sensitive feature, and five features into categorical features. For binary features, we then transformed values with a sigmoid function and binarized them at a cutoff value of 0.5. For categorical features, we rounded the initial values down to the nearest whole number.

4.2 Existing datasets

We also tested the genetic algorithm on two existing educational datasets:

(1) **MATHia** dataset includes 458 records of 49 students and 198 features of statistical information about students' actions (e.g., answers submitted, the number of attempts students made, the hints they accessed) and outcomes during the usage of MATHia, which is designed to teach math to middle school and high school students. In addition, information about students' identities were obtained through an open-ended, free response survey that allowed students to describe themselves with up to twenty statements [49]. We used the presence of gendered responses as a sensitive feature (that is, whether gender was an important part of their identity). The labels in the dataset indicate whether a student was "gaming the system", a specific type of disengagement in which students attempt to make progress with minimal learning by, for example, repeatedly guessing consecutive numbers as answers until finding the correct answer [4].
(2) **Student Performance** dataset [25] includes 395 students and their multiple-choice, person-level survey responses. The dataset contains students' demographic information and students' grades from two courses. We extracted 33 features and predicted whether final grades were above or below the median. Whether students live in urban or rural areas was used as the sensitive feature, based on previous research indicating that this is a student characteristic that relates to unfairness in machine learning [59].

5 EXPERIMENTS

5.1 Model training

We selected a simple machine learning method, logistic regression, as the model in the evaluation phase. Models were trained with 4-fold cross-validation. For each fold, we randomly selected 75% of the data for training and the other 25% of the data for testing. If an individual was represented in more than one instance (i.e., hierarchical structure in the data), we used group 4-fold cross-validation to prevent person-level data leakage. We calculated an unfairness score for each fold using the sensitive feature and used the mean unfairness across folds as the unfairness for the dataset. We set the genetic algorithm population size as 100, the number of

iterations as 50, the gene mutation rate as 0.002, and the selection rate for crossover as 0.6 (as used in many GA algorithms [43, 45]). We ran all experiments on a Macbook Pro M1 Max with 10-core CPU, 32GB RAM, and MacOS 13.5.

5.2 Effectiveness of the proposed algorithm

We conducted several experiments to assess the effectiveness of the proposed algorithm (RQ1). We simulated unfairness benchmark datasets for both real-world reference datasets and generated reference datasets, while testing three notions of unfairness to determine if our algorithm can generate datasets that fulfill various requirements. To assess the usefulness of simulated dataset, we applied the model that was trained on the simulated dataset to the original dataset and measured performance by calculating both accuracy (ACC) and area under the receiver operating characteristic curve (AUC) scores. ACC measures the number of correct predictions divided by the number of total predictions (i.e., proportion correct). AUC score measures how well a model can produce probability scores to discriminate among different classes across all possible thresholds.

5.3 Generalizability of our algorithm

For the following experiments, we used the *Simulated Dataset (3 types)* as the default reference dataset and overall accuracy equality as the unfairness metric to explore the performance of our algorithm. To assess the generalizability of our algorithm (RQ2), we examined whether the difficulty of the problems and the number of samples affect the usability of our algorithm. To test increasing difficulty of classification problems, we randomly changed a proportion of the labels in the dataset (.01, .1, .2, and .5 proportions), and simulated unfair datasets for each value. A higher proportion means there is more noise in the labels, which leads to a harder classification problem. We also varied the number of samples (500, 1000, 5000) in the dataset to identify if the algorithm can generate an unfair dataset when the dataset is larger or smaller. We additionally used random forest and extremely randomized trees [16, 38] to evaluate the unfairness of the dataset, in order to identify if the simulated dataset also presents unfairness using other classifiers.

5.4 Performance of existing bias mitigation algorithms on simulated unfairness benchmark datasets

We tested whether existing bias mitigation methods can remove unfairness from a simulated dataset under different measurements of unfairness. We specifically tested the simulated dataset (3 types). The metrics chosen to measure fairness were overall accuracy equality, statistical equality, and calibration equality. We tested reweighing [44], disparate impact remover [35], and equalized odds [42] methods for the existing fairness approach.

6 RESULTS

In this section, we describe the results of our experiments. In section 6.1, we present the results of our dataset generation algorithm when using different definitions of unfairness as well as different reference datasets. We also demonstrate the impact of different gene mutation rates on the resulting dataset. In section 6.3, we present datasets generated to reflect increasingly difficulty prediction tasks and larger datasets. In section 6.4, we present the performance of three extant fair machine learning methods on our simulated unfairness benchmark dataset.

6.1 Results of our algorithm (RQ1)

The proposed algorithm was able to create datasets encoding unfairness using both real-world and simulated datasets as the reference. Table 1 demonstrates the results. For each generated dataset, we measured unfairness using overall accuracy, statistical parity, and calibration equality. Our algorithm successfully increased unfairness under these three different statistical metrics. Overall accuracy equality, statistical parity, and calibration increased by 0.143 (± 0.018), 0.129 (± 0.015), and 0.103 (± 0.010) respectively. Overall, unfairness increased by 0.125 (± 0.022) across datasets and metrics, representing a 156.3% increase. The AUC and ACC scores, in comparison, did not change dramatically. When the model was trained on the generated dataset and tested on the original reference dataset, the AUC score was comparable to that of the model trained on the original dataset, demonstrating that the relationship to the original dataset is preserved with this method.

We additionally explored different gene mutation rates. In short, we used one dataset and unfairness metric while varying the gene mutation rate. The results show that the proposed method can generate a dataset encoding unfairness given all three different gene mutation rates. When we increased the gene mutation rate from .002 to .004, unfairness increased only slightly, by .014. However, comparing results between mutation rates of .004 and .008, the unfairness slightly decreased (by .008). This likely indicates that the larger gene mutation rate also led to too much randomness in the search process to achieve unfairness. We also found that higher mutation rates did cause slight decreases in accuracy, demonstrated by lower ACC and AUC scores for the simulated dataset as well as a lower AUC score on the original dataset.

We also experimented with changes in population size. We predicted that increased population size would affect the performance of our algorithm with respect to both unfairness and speed. A larger population size was predicted to increase unfairness in the result, given that each iteration simulated a larger number of datasets, providing a denser sampling of the search space for optimization. Our results support this intuition and are shown in their entirety in Table 2. In this experiment, the same reference dataset and unfairness metric were used while the population size was varied. As the population size increased, compared with population of 50, datasets with higher unfairness scores were generated (up to .035 higher with a population of 200). However, the elapsed time also increased (up to 2x), reflecting the increased computational task.

6.2 Analysis of generated datasets

To gain a deeper understanding of which type of feature is likely to be modified using this method, we analyzed the generated datasets to determine what had changed compared to the initial reference dataset. The last column in Table 1 demonstrates that each dataset has only been changed by around 6.2% while successfully inducing bias. We additionally investigated the correlation between the type

Table 1: Results of the modified genetic algorithm on different reference datasets, including MATHia, student performance, simulated 3-type, and simulated 1-type datasets. For each dataset, we used calibration equality, overall accuracy, and statistical parity as the metrics to identify unfairness.

Dataset		Original			Generated					Change in unfairness
	Metric	Unfairness	AUC	ACC	Unfairness	AUC	ACC	AUC on original dataset	Prop. values unchanged	
MATHia	calibration equality	.053	.869	.930	.162	.757	.897	.842	.966	.109
	overall accuracy equality	.059	.869	.930	.178	.766	.893	.873	.982	.119
	statistical parity	.066	.869	.930	.186	.852	.900	.877	.973	.120
Student performance	calibration equality	.052	.756	.698	.164	.742	.668	.752	.974	.112
	overall accuracy equality	.044	.756	.698	.182	.743	.670	.754	.980	.138
	statistical parity	.176	.756	.698	.320	.736	.675	.753	.975	.144
Simulated dataset (3 types)	calibration equality	.036	.895	.807	.132	.869	.772	.892	.961	.096
	overall accuracy equality	.040	.895	.807	.200	.863	.775	.890	.955	.160
	statistical parity	.183	.895	.807	.322	.876	.793	.891	.956	.139
Simulated dataset (1 type)	calibration equality	.042	.955	.897	.134	.940	.867	.954	.949	.092
	overall accuracy equality	.009	.955	.897	.162	.939	.844	.957	.943	.153
	statistical parity	.199	.955	.897	.311	.940	.870	.956	.952	.112
Average		.080	.869	.833	.204	.835	.802	.866	.964	.125

Table 2: Results of simulated unfair datasets using different gene mutation rates and different population sizes (50, 100, and 200).

		Original			Generated				Time
		Unfairness	AUC	ACC	Unfairness	AUC	ACC	AUC on original dataset	
Mutation rate	.002	.040	.894	.807	.200	.863	.775	.890	1x(6m10s)
	.004	.040	.894	.807	.214	.853	.763	.891	1x(6m28s)
	.008	.040	.894	.807	.206	.845	.765	.887	1x(6m6s)
Population	50	.040	.894	.807	.185	.862	.774	.892	0.5x (3m5s)
	100	.040	.894	.807	.200	.863	.775	.890	1x (6m10s)
	200	.040	.894	.807	.220	.860	.783	.889	2x (12m35s)

of feature and the amount of change. The correlation between the number of possible values and the amount of change is 0.787 on average across datasets and unfairness definitions, even the lowest correlation reaches 0.697. We found that features with a larger range of possible values experienced more change. Finally, we computed the correlation between feature importance and the amount of change but found it not to be statistically significant.

6.3 Results of generalizability (RQ2)

We simulated datasets with 500, 1000, and 5000 rows and used each of these as a reference dataset for our algorithm. All results showed increased unfairness in the generated dataset when compared to the reference. When the number of features remained constant and the number of examples increased, the total unfairness of the resulting dataset was more difficult to increase. When the reference dataset had 500 rows, the generated dataset had an unfairness of .307. When

Table 3: Results of using reference datasets with different numbers of examples (500, 1000, and 5000) but same number of features and with different difficulties (A higher flip ratio generally means a harder classification problem).

		Original			Generated				Time
		Unfairness	**AUC**	**ACC**	**Unfairness**	**AUC**	**ACC**	**AUC on original dataset**	
# of examples	500	.090	.853	.764	.307	.815	.730	.853	0.8x (5m)
	1000	.040	.894	.807	.200	.863	.775	.890	1x (6m10s)
	5000	.022	.837	.754	.087	.821	.737	.836	2.5x (15m35s)
Flip ratio	.01	.040	.894	.807	.200	.863	.775	.890	1x (6m10s)
	.1	.054	.830	.754	.200	.826	.741	.832	1x(6m11s)
	.2	.025	.834	.773	.206	.805	.738	.832	1x(6m28s)
	.5	.067	.683	.640	.228	.686	.643	.684	1x(6m7s)

Table 4: We tested all 4 datasets as reference, logistic regression (LR), random forest (RF), and extremely randomized trees (ERT) as the fitness function, and overall accuracy equality as unfairness metric. All three classifiers were trained on the simulated unfair dataset. The results shown in the table are unfairness changes.

Classifier in fitness function	Classifier in evaluation											
	Student			MATHia			Simulated (3 types)			Simulated (1 type)		
	LR	**RF**	**ERT**	**LR**	**RF**	**ERT**	**LR**	**RF**	**ERT**	**LR**	**RF**	**ERT**
LR	.138	.000	.005	.113	.007	.002	.160	.087	.051	.152	.042	.049
RF	.034	.088	.039	.005	.027	-.005	.034	.142	.052	.014	.042	.021
ERT	.028	.037	.124	.016	.015	.025	.037	.067	.137	.014	.012	.056

the reference dataset had 1000 rows, unfairness increased, but only to .200. When the reference dataset had 5000 rows, the unfairness increased even less, to .087. While unfairness did not achieve the same amount of change with larger dataset size, the model learned on the simulated unfair datasets achieved comparable AUC scores as the model trained on the original dataset – a difference of only .002 (±.002). We also calculated the time elapsed; increasing size increased the running time approximately linearly since logistic regression was used as the model. Running time is largely dominated by machine learning model training time, and thus depends on the model used in the evaluation operation. The detailed results are shown in Table 3. The results imply that the modified genetic algorithm might not be suitable for handling large datasets.

We also investigated the impact of increasingly difficult categorization problems on our algorithm and present the full results in Table 3. As the difficulty increased, the AUC and accuracy scores decreased overall. Making accurate predictions is demonstrably more difficult. However, in each case, our algorithm still successfully increased unfairness while keeping the AUC score of the model trained on the new dataset but tested on the original dataset similar.

The previously described experiments measured fitness using logistic regression. We therefore designed an additional experiment to test our generated datasets against other models. We trained a logistic regression classifier, a random forest classifier, and an extremely randomized trees classifier on one unfair dataset generated with the logistic regression fitness function. The unfairness of the logistic regression model was .200, while unfairness of the random forest and extremely randomized trees classifiers were lower, at .124 and .097, respectively. Therefore, generated datasets may not express unfairness in classifiers other than the one used to measure fitness, at least to the same extent. However, we demonstrate experimentally that our algorithm can generate unfair benchmark datasets using different models as the fitness function in Table 4.

While the modified genetic algorithm can be applied to different sizes of datasets, problems with varying difficulties, and different classifiers, we acknowledge that the algorithm only works for tabular datasets. There may be potential for the algorithm to be applied to datasets in other domains where genetic algorithms have been valuable, such as object recognition in computer vision. However, the algorithm cannot currently handle sequential datasets, which is also a common format in education, because of the complex constraints (e.g., logical and statistical dependencies between elements in a sequence) that are not yet considered during mutation and crossover in the genetic algorithm.

6.4 Fair algorithms (RQ3)

We tested different unfairness mitigation methods, including reweighing, disparate impact remover, and equalized odds algorithms on the simulated unfair dataset to determine whether unfairness mitigation approaches can improve fairness in a model trained on the generated dataset. The results are shown in Table 5. We observed that reweighing substantially decreased the statistical inequality (by .160); disparate impact remover only slightly reduced overall

Table 5: Changes in unfairness scores under different notions of fairness when applying various methods for bias mitigation.

Unfairness definition	Initial unfairness	Reweighing		Disparate impact remover		Equalized odds	
		Unfairness	Change	Unfairness	Change	Unfairness	Change
Calibration equality	.082	.078	-.004	.069	-.013	.135	+.053
Overall accuracy equality	.189	.173	-.016	.179	-.010	.228	+.039
Statistical equality	.335	.175	-.160	.325	-.010	.068	-.267

accuracy inequality (.010) and calibration equality (.013); and equalized odds substantially decreased statistical inequality (.267). This result demonstrates that existing bias mitigation methods decreased unfairness to some extent, but not necessarily equally across all definitions, illustrating the importance of exploring – and creating – datasets with different specific types of unfairness embedded in them.

7 CONCLUSION

In this paper, we presented a novel data generation method using a genetic algorithm to intentionally induce bias in datasets from educational domains for multiple statistical fairness metrics. Based on our experiments, nearly all datasets exhibit unfairness of less than 0.1. When evaluating bias mitigation methods, the improvements achieved by different algorithms may not be significant. Current benchmark datasets limit the types of unfairness that can be studied, especially in the education field. Furthermore, even if new debiasing algorithms emerge, the evaluation methods may not accurately quantify the performance of these algorithms if we continue to use a small number of outdated benchmarks. Researchers may therefore encounter growing challenges in selecting among various bias mitigation algorithms. Our modified genetic algorithm is scalable and can generate patterns of unfairness that may not be captured in the few benchmark datasets commonly examined in fairness research. By varying the parameters given to the genetic algorithm, a variety of unfair datasets can be generated with this method and can therefore be used to reflect situations with different fairness goals. It is also robust to categorization problems of varying difficulties. Though the unfairness mitigation methods tested only decreased unfairness under certain statistical definitions, this likely reflects the fact that there is no one-size-fits-all approach to measuring – much less mitigating – algorithmic unfairness [72]. Moreover, the current approach exhibits limitations when dealing with large datasets: our modified genetic algorithm may not increase unfairness for large datasets as much, and the running time increases approximately linearly. In the future, we plan to extend this method to generate additional types of datasets, such as those with sequential data, and to generate model-agnostic unfairness benchmark datasets by expanding the fitness function to consider unfairness with respect to multiple machine learning models. Ultimately, we expect that datasets generated with this method can be used for benchmarking and testing for safe future research on the measurement and mitigation of algorithmic unfairness, particularly for situations where collecting unfair real-world data may itself be a questionable, unfair action.

ACKNOWLEDGMENTS

This research was supported by NSF grant no. 2000638. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] John M. Abowd and Lars Vilhuber. 2008. How protective are synthetic data?. In *Privacy in Statistical Databases: UNESCO Chair in Data Privacy International Conference, PSD 2008, Istanbul, Turkey, September 24-26, 2008. Proceedings*. Springer, Springer, Heidelberg, 239–246.

[2] Ricardo Aguiar and MTAG Collares-Pereira. 1992. TAG: a time-dependent, autoregressive, Gaussian model for generating synthetic hourly radiation. *Solar energy* 49, 3 (1992), 167–174.

[3] Samuel A. Assefa, Danial Dervovic, Mahmoud Mahfouz, Robert E. Tillman, Prashant Reddy, and Manuela Veloso. 2021. Generating synthetic data in finance: opportunities, challenges and pitfalls. In *Proceedings of the First ACM International Conference on AI in Finance (ICAIF '20)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3383455.3422554

[4] Ryan S. Baker, Albert T. Corbett, Kenneth R. Koedinger, and Angela Z. Wagner. 2004. Off-task behavior in the cognitive tutor classroom: When students "game the system". In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04)*. ACM, New York, NY, 383–390. https://doi.org/10.1145/985692.985741

[5] Ryan S. Baker, Lief Esbenshade, Jonathan Vitale, and Shamya Karumbaiah. 2023. Using demographic data as predictor variables: A questionable choice. *Journal of Educational Data Mining* 15, 2 (June 2023), 22–52. https://doi.org/10.5281/zenodo.7702628 Number: 2.

[6] Ryan S. Baker and Aaron Hawn. 2021. Algorithmic bias in education. *International Journal of Artificial Intelligence in Education* 32 (Nov. 2021), 1052–1092. https://doi.org/10.1007/s40593-021-00285-9

[7] Michelle Bao, Angela Zhou, Samantha Zottola, Brian Brubach, Brian Brubach, Sarah Desmarais, Aaron Horowitz, Kristian Lum, and Suresh Venkatasubramanian. 2021. It's COMPASlicated: The messy relationship between RAI datasets and algorithmic fairness benchmarks. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* 1 (Dec. 2021), 1–18. https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/92cc227532d17e56e07902b254dfad10-Abstract-round1.html

[8] Enrico Barbierato, Marco L. Della Vedova, Daniele Tessera, Daniele Toti, and Nicola Vanoli. 2022. A methodology for controlling bias and fairness in synthetic data generation. *Applied Sciences* 12, 9 (Jan. 2022), 4619. https://doi.org/10.3390/app12094619 Number: 9 Publisher: Multidisciplinary Digital Publishing Institute.

[9] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org. http://www.fairmlbook.org.

[10] Solon Barocas and Andrew D. Selbst. 2016. Big data's disparate impact. *California Law Review* 104, 671 (2016), 671–732.

[11] Clara Belitz, Lan Jiang, and Nigel Bosch. 2021. Automating procedurally fair feature selection in machine learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. Association for Computing Machinery, New York, NY, 379–389. https://doi.org/10.1145/3461702.3462585

[12] Clara Belitz, Jaclyn Ocumpaugh, Steven Ritter, Ryan S. Baker, Stephen E. Fancsali, and Nigel Bosch. 2022. Constructing categories: Moving beyond protected classes in algorithmic fairness. *Journal of the Association for Information Science and Technology* (2022), 1–6. https://doi.org/10.1002/asi.24643

[13] 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.

[14] Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness

benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 1004–1015. https://doi.org/10.18653/v1/2021.acl-long.81

[15] Claire McKay Bowen and Fang Liu. 2020. Comparative study of differentially private data synthesis methods. *Statist. Sci.* 35, 2 (May 2020), 280–307. https://doi.org/10.1214/19-STS742 Publisher: Institute of Mathematical Statistics.

[16] Leo Breiman. 2001. Random forests. *Machine Learning* 45, 1 (Oct. 2001), 5–32. https://doi.org/10.1023/A:1010933404324

[17] Brent Bridgeman. 2013. 13 Human Ratings and Automated Essay Evaluation. *Handbook of automated essay evaluation: Current applications and new directions* (2013), 221.

[18] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: Experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*. 108–122.

[19] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. PMLR, New York, NY, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html ISSN: 2640-3498.

[20] Gregory Caiola and Jerome P. Reiter. 2010. Random forests for generating partially synthetic, categorical data. *Trans. Data Priv.* 3, 1 (2010), 27–42.

[21] Alessandro Castelnovo, Riccardo Crupi, Nicole Inverardi, Daniele Regoli, and Andrea Cosentini. 2022. Investigating bias with a synthetic data generator: Empirical evidence and philosophical interpretation. http://arxiv.org/abs/2209.05889 arXiv:2209.05889 [cs, stat].

[22] Richard J. Chen, Ming Y. Lu, Tiffany Y. Chen, Drew F. K. Williamson, and Faisal Mahmood. 2021. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering* 5, 6 (June 2021), 493–497. https://doi.org/10.1038/s41551-021-00751-8

[23] Chanachok Chokwitthaya, Yimin Zhu, Supratik Mukhopadhyay, and Amirhosein Jafari. 2020. Applying the Gaussian mixture model to generate large synthetic data from a small data set. In *Construction Research Congress 2020: Computer Applications*. American Society of Civil Engineers, Tempe, Arizona, 1251–1260.

[24] Jade Mai Cock, Muhammad Bilal, Richard Davis, Mirko Marras, and Tanja Kaser. 2023. Protected attributes tell us who, behavior tells us how: A comparison of demographic and behavioral oversampling for fair student success modeling. In *LAK23: 13th International Learning Analytics and Knowledge Conference*. ACM, Arlington TX USA, 488–498. https://doi.org/10.1145/3576050.3576149

[25] Paulo Cortez and Alice Maria Gonçalves Silva. 2008. Using data mining to predict secondary school student performance. In *Proceedings of 5th Future Business Technology Conference*. EUROSIS-ETI, Porto, Portugal, 5–12.

[26] Sasha Costanza-Chock, Inioluwa Deborah Raji, and Joy Buolamwini. 2022. Who audits the auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 1571–1583. https://doi.org/10.1145/3531146.3533213

[27] Charles Darwin. 2004. *On the Origin of Species, 1859*. Routledge.

[28] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. 2021. Retiring adult: New datasets for fair machine learning. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., New York, 6478–6490. https://proceedings.neurips.cc/paper/2021/file/32e54441e6382a7fbacbbbaf3c450059-Paper.pdf

[29] Jörg Drechsler. 2010. Using support vector machines for generating synthetic datasets. In *International Conference on Privacy in Statistical Databases*. Springer, Springer, Berlin, Heidelberg, 148–161. https://doi.org/10.1007/978-3-642-15838-4_14

[30] Jörg Drechsler. 2011. *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation*. Vol. 201. Springer Science & Business Media, New York, NY. https://doi.org/10.1007/978-1-4614-0326-5

[31] Cynthia Dwork. 2010. Differential privacy in new settings. In *Proceedings of the 2010 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA) (Proceedings)*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 174–183. https://doi.org/10.1137/1.9781611973075.16

[32] Cristóbal Esteban, Stephanie L. Hyland, and Gunnar Rätsch. 2017. Real-valued (medical) time series generation with recurrent conditional gans. *arXiv preprint arXiv:1706.02633* (2017), 13.

[33] Virginia Eubanks. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Publishing Group, New York, NY, USA.

[34] Alessandro Fabris, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. 2022. Algorithmic fairness datasets: the story so far. *Data Mining and Knowledge Discovery* 36, 6 (Nov. 2022), 2074–2152. https://doi.org/10.1007/s10618-022-00854-z

[35] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, New York, NY, USA, 259–268. https://doi.org/10.1145/2783258.2783311

[36] Gianni Fenu, Roberta Galici, and Mirko Marras. 2022. Experts' view on challenges and needs for fairness in artificial intelligence for education. In *International Conference on Artificial Intelligence in Education*. Springer, 243–255.

[37] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 2015. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security* (Denver, Colorado, USA). Association for Computing Machinery, New York, NY, USA, 1322–1333.

[38] Pierre Geurts, Damien Ernst, and Louis Wehenkel. 2006. Extremely randomized trees. *Machine Learning* 63, 1 (April 2006), 3–42. https://doi.org/10.1007/s10994-006-6226-1

[39] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM* 63, 11 (2020), 139–144.

[40] Aman Gupta, Deepak Bhatt, and Anubha Pandey. 2021. Transitioning from real to synthetic data: Quantifying the bias in model. arXiv. http://arxiv.org/abs/2105.04144 arXiv:2105.04144 [cs].

[41] Inken Hagestedt, Yang Zhang, Mathias Humbert, Pascal Berrang, Tang Haixu, Wang XiaoFeng, and Michael Backes. 2019. MBeacon: Privacy-preserving beacons for DNA methylation data. In *Network and Distributed Systems Security (NDSS) Symposium 2019*. Network and Distributed Systems Security (NDSS) Symposium, San Diego, CA, USA, 15. https://dx.doi.org/10.14722/ndss.2019.23064

[42] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016).

[43] K.A. De Jong. 1975. *An Analysis of the Behavior of a Class of Genetic Adaptive Systems*. Ph. D. Dissertation. University of Michigan, Ann Arbor, MI, USA.

[44] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33, 1 (Oct. 2012), 1–33. https://doi.org/10.1007/s10115-011-0463-8

[45] Sourabh Katoch, Sumit Singh Chauhan, and Vijay Kumar. 2021. A review on genetic algorithm: past, present, and future. *Multimedia Tools and Applications* 80 (2021), 8091–8126.

[46] Os Keyes. 2018. The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (Nov. 2018), 88:1–88:22. https://doi.org/10.1145/3274357

[47] Huda Khayrallah and Philipp Koehn. 2018. On the Impact of Various Types of Noise on Neural Machine Translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*. Association for Computational Linguistics, Melbourne, Australia, 74–83. https://doi.org/10.18653/v1/W18-2709

[48] Adam Kortylewski, Bernhard Egger, Andreas Schneider, Thomas Gerig, Andreas Morel-Forster, and Thomas Vetter. 2019. Analyzing and reducing the damage of dataset bias to face recognition With synthetic data. https://openaccess.thecvf.com/content_CVPRW_2019/html/BEFA/Kortylewski_Analyzing_and_Reducing_the_Damage_of_Dataset_Bias_to_Face_CVPRW_2019_paper.html

[49] Manford H. Kuhn and Thomas S. McPartland. 2017. An empirical investigation of self-attitudes. In *Sociological Methods*. Routledge, England, UK, 167–182.

[50] 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 and Knowledge Discovery* 12, 3 (2022), e1452. https://doi.org/10.1002/widm.1452

[51] Adam Lipowski and Dorota Lipowska. 2012. Roulette-wheel selection via stochastic acceptance. *Physica A: Statistical Mechanics and its Applications* 391, 6 (2012), 2193–2196.

[52] Qun Liu, Supratik Mukhopadhyay, Yimin Zhu, Ravindra Gudishala, Sanaz Saeidi, and Alimire Nabijiang. 2019. Improving route choice models by incorporating contextual factors via knowledge distillation. In *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, IEEE, Budapest, Hungary, 1–8. https://doi.org/10.1109/IJCNN.2019.8852482

[53] Michael Madaio, Su Lin Blodgett, Elijah Mayfield, and Ezekiel Dixon-Román. 2021. Beyond "fairness:" Structural (in)justice lenses on AI for education. In *The Ethics of Artificial Intelligence in Education: Current Challenges, Practices and Debates*, W. Holmesand and K. Porayska-Pomsta (Eds.). Routledge, England, UK, 24. http://arxiv.org/abs/2105.08847 arXiv: 2105.08847.

[54] Daniel McDuff, Shuang Ma, Yale Song, and Ashish Kapoor. 2019. Characterizing Bias in Classifiers using Generative Models. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc., New York, 1–12. https://proceedings.neurips.cc/paper/2019/file/7f018eb7b301a66658931cb8a93fd6e8-Paper.pdf

[55] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *Comput. Surveys* 54, 6 (July 2021), 115:1–115:35. https://doi.org/10.1145/3457607

[56] Oren Melamud and Chaitanya Shivade. 2019. Towards Automatic Generation of Shareable Synthetic Clinical Notes Using Neural Language Models. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics, Minneapolis, Minnesota, USA, 35–45.

https://doi.org/10.18653/v1/W19-1905
[57] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. 2019. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE symposium on security and privacy (SP)*. IEEE, IEEE, San Fransisco, CA, US, 691–706. https://doi.org/10.1109/SP.2019.00029
[58] Seyedali Mirjalili and Seyedali Mirjalili. 2019. Genetic algorithm. *Evolutionary Algorithms and Neural Networks: Theory and Applications* 780 (2019), 43–55. https://doi.org/10.1007/978-3-319-93025-1_4
[59] Jaclyn Ocumpaugh, Ryan Baker, Sujith Gowda, Neil Heffernan, and Cristina Heffernan. 2014. Population validity for Educational Data Mining models: A case study in affect detection. *British Journal of Educational Technology* 45, 3 (2014), 487–501.
[60] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns* 2, 11 (Nov. 2021), 100336. https://doi.org/10.1016/j.patter.2021.100336
[61] Kenny Peng, Arunesh Mathur, and Arvind Narayanan. 2021. Mitigating dataset harms requires stewardship: Lessons from 1000 papers. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* 1 (2021), 16.
[62] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. 2017. On fairness and calibration. *Advances in neural information processing systems* 30 (2017).
[63] Apostolos Pyrgelis, Carmela Troncoso, and Emiliano De Cristofaro. 2017. Knock knock, who's there? Membership inference on aggregate location data. In *Network and Distributed Systems Security (NDSS) Symposium 2018*. Network and Distributed Systems Security (NDSS) Symposium, San Diego, CA, USA, 15. http://dx.doi.org/10.14722/ndss.2018.23183
[64] Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. AI and the everything in the whole wide world benchmark. In *35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks*. Neural Information Processing Systems, Virtual.
[65] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 1668–1678. https://doi.org/10.18653/v1/P19-1163
[66] Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R. Brubaker. 2020. How we've taught algorithms to see identity: Constructing race and gender in image databases for facial analysis. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (May 2020), 1–35. https://doi.org/10.1145/3392866
[67] Daniel Schiff. 2021. Out of the laboratory and into the classroom: The future of artificial intelligence in education. *AI & Society* 36, 1 (March 2021), 331–348. https://doi.org/10.1007/s00146-020-01033-8
[68] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, San Jose, CA, 3–18.
[69] Ravindra Singh, Bikash C Pal, and Rabih A Jabr. 2009. Statistical representation of distribution system loads using Gaussian mixture model. *IEEE Transactions on Power Systems* 25, 1 (2009), 29–37.
[70] Frank Stinar and Nigel Bosch. 2022. Algorithmic unfairness mitigation in student models: When fairer methods lead to unintended results. In *Proceedings of the 15th International Conference on Educational Data Mining*. International Educational Data Mining Society, Durham, UK, 606–611. https://doi.org/10.5281/zenodo.6853135
[71] Boris van Breugel, Trent Kyono, Jeroen Berrevoets, and Mihaela van der Schaar. 2021. DECAF: Generating fair synthetic data using causally-aware generative networks. In *Advances in Neural Information Processing Systems*, Vol. 34. Curran Associates, Inc., 22221–22233. https://proceedings.neurips.cc/paper/2021/hash/ba9fab001f67381e56e410575874d967-Abstract.html
[72] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*. Association for Computing Machinery, New York, NY, 1–7.
[73] Angelina Wang, Vikram V Ramaswamy, and Olga Russakovsky. 2022. Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 336–349. https://doi.org/10.1145/3531146.3533101
[74] Darrell Whitley. 1994. A genetic algorithm tutorial. *Statistics and computing* 4, 2 (1994), 65–85.
[75] Linda F. Wightman. 1998. LSAC National Longitudinal Bar Passage Study. LSAC Research Report Series. (1998).
[76] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. 2018. FairGAN: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, Seattle, WA, 570–575. https://doi.org/10.1109/BigData.2018.8622525
[77] Jinsung Yoon, James Jordon, and Mihaela Van Der Schaar. 2018. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*. ICLR, Vancouver, Canada, 11.
[78] Jie Yu, Jaume Amores, Nicu Sebe, and Qi Tian. 2006. A new study on distance metrics as similarity measurement. In *2006 IEEE International Conference on Multimedia and Expo*. IEEE, IEEE, Toronto, ON, Canada, 533–536.
[79] Jun Zhang, Graham Cormode, Cecilia M Procopiuc, Divesh Srivastava, and Xiaokui Xiao. 2017. Privbayes: Private data release via Bayesian networks. *ACM Transactions on Database Systems (TODS)* 42, 4 (2017), 1–41.
[80] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. 2021. A comprehensive survey on transfer learning. *Proc. IEEE* 109, 1 (Jan. 2021), 43–76. https://doi.org/10.1109/JPROC.2020.3004555