UNDER REVIEW

RESCIENCE C

Replication / ML Reproducibility Challenge 2022

On the reproducibility of "Fairness Guarantees under Demographic Shift"

Anonymous[1, ID]

[1] Anonymous Institution

03 February 2023

Edited by
(Editor)

Reviewed by
(Reviewer 1)
(Reviewer 2)

Received
03 February 2023

Published
–

DOI
–

Reproducibility Summary

Scope of Reproducibility – The original authors' [1] main contribution is the family of *Shifty* algorithms, which can guarantee that certain fairness constraints will hold with high confidence even after a demographic shift in the deployment population occurs. They claim that Shifty provides these high-confidence fairness guarantees without a loss in model performance, given enough training data.

Methodology – The code provided by the original paper was used, and only some small adjustments needed to be made in order to reproduce the experiments. All model specifications and hyperparameters from the original implementation were used. Extending beyond reproducing the original paper, we investigated the sensibility of *Shifty* to the size of the bounding intervals limiting the possible demographic shift, and ran shifty with an additional optimization method.

Results – Our results approached the results reported in the original paper. They supported the claim that *Shifty* reliably guarantees fairness under demographic shift, but could not verify that *Shifty* performs at no loss of accuracy.

What was easy – The theoretical framework laid out in the original paper was well explained and supported by additional formulas and proofs in the appendix. Further, the authors provided clear instructions on how to run the experiments and provided necessary hyperparameters.

What was difficult – While an open-source implementation of *Shifty* was provided and was debugged with relatively low time investment, the code did not contain extensive documentation and was complex to understand. It was therefore difficult to verify that each part of the code functions as expected and to expand upon the existing experiments. Further, certain hyperparameter and model specifications deviated between the provided code and the original paper, which made it challenging to know which specifications to apply when reproducing.

Communication with original authors – The first author of the original paper was contacted, but unfortunately we have yet to receive a reply.

Copyright © 2023 Anonymous, released under a Creative Commons Attribution 4.0 International license.
Correspondence should be addressed to ()
The authors have declared that no competing interests exists.
Code is available at https://anonymous.4open.science/r/fairness_guarantees.

1 Introduction

Machine Learning models have shown great performance in multiple domains over the last decade. However, recent investigations into different real-world applications have shown that some models have bias in them and thus make unfair decisions [2]. That is why algorithms have been created which provide high-confidence fairness guarantees [3] [4]. Fairness in the context of decision-making is the absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics, e.g. race or sex [2]. However, current models don't take the potential probabilistic increase or decrease of subgroups in the deployment data into account. We call this phenomenon a *demographic shift*. A demographic shift is a distribution change between the training and deployment distributions, which can be explained by a shift in the marginal distribution of a single random variable. This makes it not possible to guarantee high-confidence fairness for the current models.

To address this issue, an algorithm has been created that provides the high-confidence guarantees that one or more user-specified fairness constraints will hold, despite a demographic shift between training and deployment without using data from the deployment environment [1].

2 Scope of reproducibility

The main contribution of the original paper [1] is *Shifty*, an algorithm which guarantees with high confidence that certain fairness constraints will hold even after a demographic shift occurs. This tackles the problem that the fairness guarantees of existing algorithms [4, 3] do not hold when a demographic shift occurs upon deployment. *Shifty* is designed to work in the following two scenarios:

1. The demographic proportions in the deployment population are known (known demographic shift).
2. The demographic proportions in the deployment population are bounded to known intervals (unknown demographic shift).

Shifty is widely applicable since it is model-agnostic, and accompanied by an open source-framework. When using *Shifty*, the user can specify a fairness attribute, for which a certain fairness constraint should hold, and a demographic attribute, which may undergo demographic shift upon deployment. The following claims about *Shifty* are made:

1. *Shifty* provides high-confidence fairness guarantees under demographic shifts.
2. Given that sufficient training data exists, *Shifty* shows no loss in accuracy compared to other models.
3. If too little training data exists, or if the fairness constraints are not satisfied, *Shifty* returns *NSF* (No Solution Found).
4. *Shifty* is model-agnostic and can be used in combination with any classification model

The goal of this paper is to test whether the claims made in the paper are reproducible, check whether *Shifty* can be assumed to work in other domains and models as well, and to verify the reproducibility of the open-source implementation of the algorithm.

3 Methodology

The authors made their code repository publicly available on GitHub [5]. Unfortunately, the code from the repository produced numerous errors which we had to resolve in order to run the experiments. Furthermore, we contributed by adding documentation, a complete environment file, result data conversion to the JSON format, and additional code to create figures, tables, and conduct statistical tests.

3.1 Shifty algorithm

The *Shifty* algorithm consists of three main parts: data partitioning, candidate selection and a fairness test.

Data partitioning Initially, the data is split in two parts, one for candidate selection (D_c) and for for the fairness test (D_f). If the same dataset were to be used for both selecting a model and performing the fairness test, then the outcome of these two steps would be correlated. By ensuring that candidate selection and the fairness test use independent sets of observations this correlation is eliminated. This is necessary to guarantee that the overall algorithm satisfies property **A** (1) below.

Candidate selection This part involves training of the candidate model. Since this is not responsible for establishing the fairness guarantees, any existing classification algorithm can be used. As in the original paper, we use a linear classification model consisting of one linear layer with no bias term. The output which are produced by this linear layer are then passed through a sign function, which returns -1 for negative signs and 1 for positive signs.

Fairness test Given the candidate model from the previous step, *Shifty* performs a fairness test by computing a high-confidence upper bound for every given fairness definition. If the upper bounds for the respective fairness definitions are all below zero, *Shifty* returns the candidate model. Otherwise, *No Solution Found (NSF)* is returned. The fairness test is based on one or multiple fairness definitions g, for which it holds that $g(\theta) > 0$ if and only if θ behaves unfairly. If the candidate model is returned, it satisfies properties **A** and **B** for all fairness definitions g shown in Equation 1 below. A model is considered fair with high confidence if property A is met. Furthermore, g' represent the fairness definitions after a demographic shift, used to evaluate Property **B**.

$$\text{A) } Pr(g(a(D)) \leq 0) \geq 1 - \delta \qquad \text{B) } Pr(g'(a(D)) \leq 0) \geq 1 - \delta \tag{1}$$

3.2 Other fairness algorithms

Shifty is compared to three families of existing fairness algorithms: Fairness Constraints, Seldonian algorithms and Fairlearn, and finally RFLearn. Fairness Constraints provides fairness without guarantees, as opposed to Seldonian algorithms and Fairlearn which do provide high-confidence guarantees. Furthermore, RFLearn promotes fair outcomes under covariate shift without guarantees. *Shifty* combines both, such that it provides fairness under demographic shift with a high-confidence guarantee. Every fairness algorithm has an implementation with a particular underlying classification and optimization method, which are summed up in Table 1.
Interestingly, the Covariance Matrix Adaptation Evolution Strategy (*CMA-ES*) is used for the optimization of the Seldonian algorithms and *Shifty*, instead of the widely used and more efficient SGD. The reason that the authors opted out of using SGD might have to do with the fact that the fairness constraints are included in the loss functions for these

algorithms. As a result, the loss function may be non-differentiable or difficult to differentiate. Therefore, CMA-ES is used since it is a derivative-free numerical optimization algorithm for non-convex problems. CMA-ES requires a lot more computations than SDG, so it will take significantly more time to train a model using CMA-ES as opposed to using SDG.

	Classification	Activation	Optimization
Seldonian	1 linear layer, no bias	sign function	SLSQP + CMA-ES
FairConst	1 linear layer, no bias	sign function	SLSQP
Fairlearn	linear SVC[1]	n/a	expgrad[2]
RFLearn	1 linear layer with bias	softmax function	SGD[3]
Shifty	1 linear layer, no bias	sign function	SLSQP + CMA-ES

[1] Support Vector Classifier
[2] exponentiated gradient reduction
[3] Stochastic Gradient Descent

Table 1. Overview of fairness models based on their classification, activation and optimization methods.

3.3 Datasets

Two datasets were used for the experiments: The UCI Adult Census (Adult) dataset [6] and the UFRGS Entrance Exam and GPA (Brazil) dataset [7].

Adult dataset: The Adult dataset includes various features and protected characteristic features like race and sex of 48,842 individuals taken during the 1994 US census. Just like the authors, we also considered a subset of the dataset corresponding to black or white individuals for running the experiments. The fairness attribute for this dataset is the race of an individual and the demographic attribute is the sex of an individual. The task used in this paper is to classify whether somebody's income is above $50,000 a year.
Brazil dataset: The Brazil dataset describes academic records for 43,303 students from a university in Brazil. For each student, the dataset includes a vector of entrance exam scores, their GPA, between 0 and 4, and the student's race and sex. The fairness attribute for this dataset is the sex of an individual and the demographic attribute is the race of an individual. The task is to classify whether a student will have a GPA above 3, given their entrance exams scores.

For both classification tasks, the fairness and the demographic attribute are not taken into account when generating predictions, but solely to evaluate the models fairness. The original paper mentioned that the data was split evenly between D_c and D_f. However in the code provided this split is 60% - 40%. For the experiments we stick to the values given in the code.

3.4 Hyperparameters

For the experiments we used the hyperparameters provided by the authors in both the original paper and the code repository. However, certain hyperparameters found in the implementation did not match with what was stated in the paper, or were not given in the paper. In these cases we used the values provided in the original implementation.

3.5 Experimental setup and code

In order to run the original author's code, we needed to make some adjustments to the codebase. Our version of the repository is available at https://anonymous.4open.science/r/

fairness_guarantees and includes instructions on how to reproduce the experiments with the provided shell scripts or batch files.

To compare *Shifty* with the other fairness algorithms, we trained every algorithm on both datasets with each of the following fairness definitions: disparate impact, demographic parity, equalized odds, predictive equality and equal opportunity. The formal definitions of these fairness definitions can be found in Appendix D2 of the original paper [1]. To account for stochastic properties during the training process, we ran each experiment for the same number of trials as was specified in the original authors' implementation, which was either 25, 20, or 10 (specified in the shell script). Further, each experiment was run of sample sizes ranging between 10k and 60k, to investigate the algorithms sensibility to dataset size. Performance and fairness of the models was compared by measuring their accuracy (*Acc*) on the classification task, their proportion *NSF*, and their *Failure Rate (FR)*. The *proportion NSF* represents the proportion of trials which did not return a solution, while *FR* is the proportion of solutions which violated property $\mathbf{A}$ (original distribution) or property $\mathbf{B}$ (deployment distribution). The original authors specified the failure rate to be equal to 0 when no solutions are found, but we changed this to *n/a*, since stating that the failure rate is equal to 0 when no solutions around in all cases does not represent the data well.

In addition to collecting and visualizing the experimental results, we performed two-sample t-tests on all trials which returned a solution, comparing the accuracy of *Shifty* to the accuracy of the best model of the respective experimental run. This was done to verify the original author's claim that Shifty performs at no loss of accuracy. Statistical significance was evaluated using a significance level $alpha = 0.05$ and as one comparison was made for each combination of fairness definition, dataset, and known/unknown bounds, Bonferroni correction was applied to prevent a family-wise increase in Type-I error.

Going beyond reproducing the experiments of the original paper, we investigated *Shifty's* sensibility to the size of the bounds limiting the unknown demographic shift. This was done by running *Shifty* for multiple bound sizes on both classification tasks and for a dataset size of 60k samples. Further, to verify *Shifty's* usability with other optimizers than CMA-ES, we also ran experiments using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm for optimization. This optimization method was already implemented by the original author, even though not being reported on in the original paper. Due to extensive running time we only applied the fairness definitions disparate impact and demographic parity for both of these extensions.

3.6 Computational requirements

CPU	RAM	% of trials	Runtime (h)
Apple M1, 8 cores	32 GB	18.75%	21
Intel i7-11850H, 8 cores	16 GB	31.25%	21
Intel i9-12900KF, 8 cores	32 GB	25%	34
Intel Xeon Gold 5118, 12 cores	200 GB	25%	13

Table 2. Overview of machines used to run the experiments and their corresponding runtimes. The percentage of trials only serves as an indication for the relative usage of each machine; the trials vary in datasets and parameters, so they are not identical.

In order to run all the experiments from the original paper [1], we divided the workload across multiple machines. Table 2 gives an overview of the machines we have used and

their corresponding runtimes. In total, it took us 89 hours of computation time to run all the experiments.

4 Results

4.1 Results reproducing original paper

Our results very much resemble the results achieved by the original author, while our statistical analysis highlights that in most cases, *Shifty* performs significantly worse than the best model on the respective classification task. Tables 3 and 4, as well as Figures 2 and 1, show the results for both datasets using disparate impact as a fairness definition, while additional results with different fairness definitions can be found in A. Over all experiments, the proportion *NSF* returned by *Shifty* ranges between 0.125 and 1. ΔAcc refers to the difference in classification accuracy between using a dataset of 10k and 60k samples. If no single solution could be found for 10k samples, this value is n/a. It is relevant to note that *Shifty's* FR is always 0, while this is not the case for the comparison models.

	Known DS				Unknown DS			
	NSF	Acc	FR	Δ Acc	NSF	Acc	FR	ΔAcc
FairConst	n/a	0.782	1.000	-0.004	n/a	0.802	1.000	-0.009
RFLearn	n/a	**0.787**	1.000	0.000	n/a	0.823	1.000	0.005
Fairlearn	n/a	0.781	1.000	-0.001	n/a	**0.842**	1.000	0.007
Quasi-SC	0.520	0.762	0.417	0.111	0.600	0.767	0.500	0.139
Shifty	0.720	0.750[1]	**0.000**	0.074	0.400	0.750[2]	**0.000**	0.167
SC	0.680	0.759	**0.000**	0.105	0.500	0.781	**0.000**	0.140

[1] significantly worse than best model, $p < 0.001$, $t = 12.987$, $df = 30$
[2] significantly worse than best model, $p < 0.001$, $t = 32.561$, $df = 24$

Table 3. Comparison between the models for the greatest sample size using the Disparate Impact fairness definition and Adult dataset. Shifty differs significantly from the best performing model, and fewer solutions are found. For the first three models NSF is not applicable, since there is always a solution returned. *DS* = Demographic Shift, *NSF* = No Solution Found, *Acc* = Accuracy, *FR* = Failure Rate, Δ *Acc* = Difference in accuracy between the largest and smallest sample size.

	Known DS				Unknown DS			
	NSF	Acc	FR	Δ Acc	NSF	Acc	FR	Δ Acc
FairConst	n/a	0.612	1.000	-0.001	n/a	0.612	1.000	0.001
RFLearn	n/a	0.648	0.040	-0.001	n/a	0.642	**0.000**	-0.004
Fairlearn	n/a	0.650	0.600	-0.001	n/a	0.649	0.600	0.000
Quasi-SC	0.000	**0.657**	0.400	0.008	0.000	**0.656**	0.200	0.007
Shifty	0.600	0.606[1]	**0.000**	0.130	0.450	0.480[2]	**0.000**	0.030
SC	0.000	0.657	0.240	0.010	0.050	0.654	**0.000**	0.009

[1] significantly worse than best model, $p < 0.001$, $t = 5.357$, $df = 33$
[2] significantly worse than best model, $p < 0.001$, $t = 22.643$, $df = 29$

Table 4. Comparison between the models for the greatest sample size using the Disparate Impact fairness definition and Brazil dataset. Shifty differs significantly from the best performing model, and fewer solutions are found.

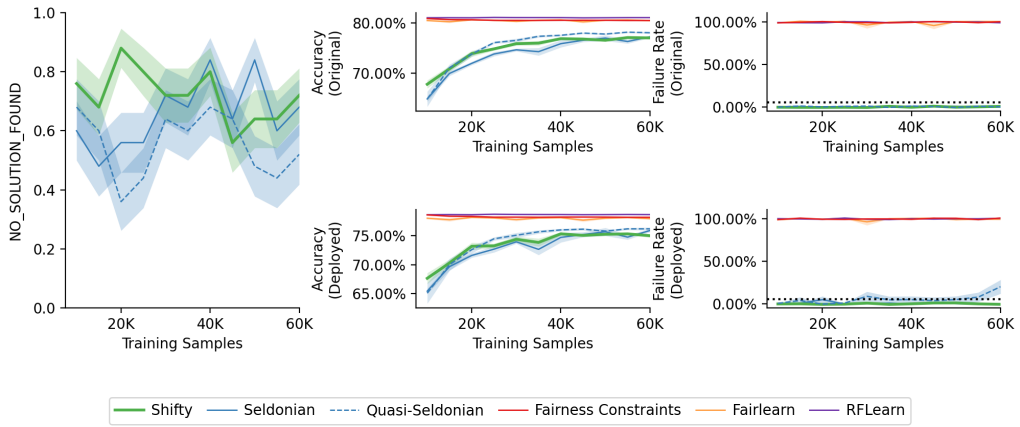

Figure 1. The results when enforcing disparate impact constraints on the UCI Adult Census dataset under known demographic shift

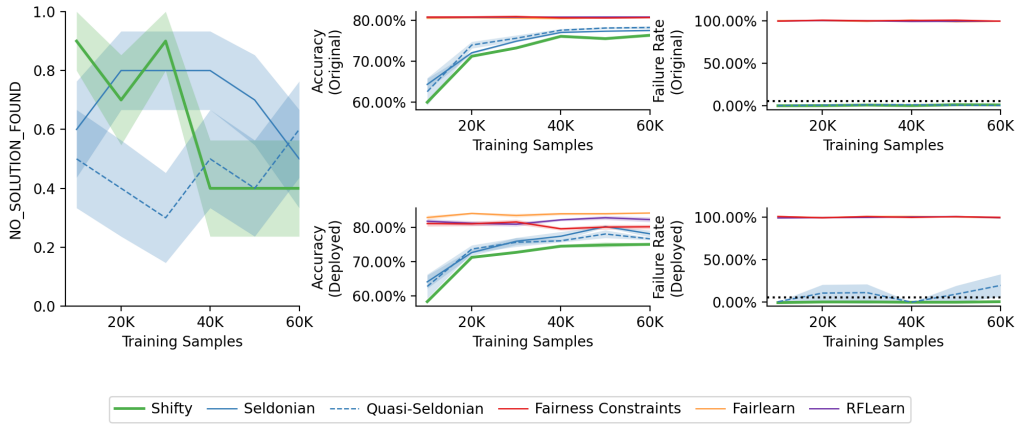

Figure 2. The results when enforcing disparate impact constraints on the UCI Adult Census dataset under unknown demographic shift

4.2 Results beyond original paper

Bound size of unknown demographic shift — Figure 3 displays how *Shifty's* performance and proportion *NSF* change when the size of the bounds defining the possible demographic shift change. A general trend can be seen that as the bounds become larger, classification accuracy decreases and the proportion *NSF* increases.

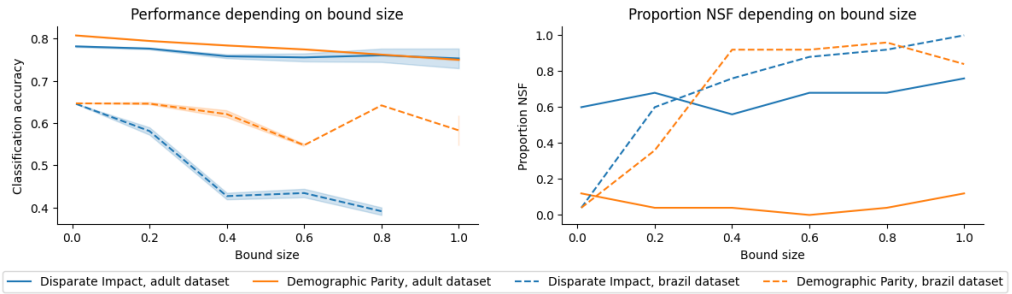

Figure 3. *Shifty's* performance and proportion NSF when demographic shift is limited to bounds of certain size. The shaded area represents the standard error.

Comparing different optimizers with Shifty — We have used three different optimization methods using the *Shifty* implementation: CMA-ES (default), BFGS and linear-shatter. Linear-shatter turned out to be impractical, because the computation time was too high. Therefore, a comparison between only CMA-ES and BFGS is shown in Table 5. We notice that the accuracy is slightly worse overall for BFGS. Furthermore, we have measured that BFGS leads to a runtime which is roughly 2x the runtime of CMA-ES. Additional results for this experiment can be found in Appendix A.3.

	Known DS				Unknown DS			
	NSF	Acc	FR	Δ Acc	NSF	Acc	FR	Δ Acc
CMA-ES	0.720	**0.750**	**0.000**	0.074	0.400	**0.750**	**0.000**	0.167
BFGS	0.440	0.724	**0.000**	0.073	0.600	0.725	**0.000**	0.210

Table 5. Comparison between CMA-ES and BFGS optimization methods for the *Shifty* implementation using the UCI Adult Census dataset and the Disparate Impact fairness definition.

5 Discussion

The reproduced results support the claims made by the authors to a large extent: (1) By design, *Shifty* guarantees that the fairness guarantees even hold under demographic shift, which the results confirm as *Shifty* never violated these constraints upon deployment. (2) The claim that, given sufficient data, *Shifty* performs at no loss of accuracy compared to other models does not hold in all problem settings. In almost all cases in which *Shifty* returned a sufficient number of solutions to conduct statistical analysis, it performed significantly worse than the best model, and when applying disparate impact as a fairness definition on the Brazil dataset, *Shifty* even performed worse than random with an accuracy of 0.48 for the unknown demographic shift setting and with the largest dataset size. As for some problem settings *Shifty's* accuracy is comparable to the other models, it would be necessary to evaluate on a case-by-case basis if the trade-off between loss of accuracy and fairness guarantees is worth it. (3) As again guaranteed by design, *Shifty* reliably returns *NSF* when it cannot guarantee the specified constraints with the data given. However, the results show that this happens more often than claimed, and sometimes even up to 100% of the cases. Our additional experiments showed that this proportion *NSF* is dependent on the size of the bounds limiting the unknown demographic shift, and it would be a valuable future research question to investigate how other hyper-parameters influence the proportion *NSF*. (4) While *Shifty* is model-agnostic by design, the original experiments did not verify this empirically as they were all run using CMA-ES. By comparing CMA-ES with BFGS, we have verified that both optimizers give comparable results, with BFGS being consistently worse than CMA-ES.
Unfortunately, the validity of the claims made is weakened by a few experimental aspects. Different methods were used to train the model parameters of *Shifty* and the comparison models, using different amounts of computational resources and model structures. As these aspects can account for differences in performance, it makes it difficult to draw final conclusions about *Shifty*. For example, Table 1 shows that the fairness models have differences in the underlying classification and optimization methods, which can impact the final accuracy of the model.

Given that the claims would hold, another limitation of *Shifty* may be its practicality when being applied to real-world problems. Since it already requires 60,000 training examples to approach competitive performance when only training a one-layer model, it may not always be feasible to obtain enough labelled data to train more complex models. The relation between model performance, model complexity, and dataset size for *Shifty* may be a valuable future research question. Further, while *Shifty* in theory works

with any classification model, it may not often find a solution when not incorporating the model's fairness into the loss function. However, no implementation of *Shifty* for popular optimization methods like Stochastic Gradient Descent (SGD) exists yet.

5.1 What was easy

The provided theory for the *Shifty* algorithm, along with the proof was helpful in understanding the algorithm thoroughly. The original paper provided a lot of additional information in the appendix for better understanding. This was helpful in not only better understanding the code, but also to verify certain functions in the implementation.

Once the code was functional, the process of running the code and getting the results was straightforward. The authors provided a batch file with multiple command lines to run the various experiments. With this batch file the necessary functions along with the given hyperparameters were successfully run after only a few adjustments.

5.2 What was difficult

Understanding the implementation: During the project, we experienced difficulties understanding the implementation provided by the authors. The open-source code implementation for the *Shifty* algorithm and the accompanying experiments was of significant size. The codebase consisted of about 15,000 lines of Python code lacked documentation, a large part of it was not relevant to the published paper, and a lack of structure made it difficult to extract the relevant material. What additionally contributed to the size and complexity of the codebase was that many code snippets were present at multiple locations in parts of the code. Fortunately, this redundancy can be fixed by defining generalized functions. Unfortunately, this was not a feasible task to accomplish during this project due to the limited time we had and the size of the codebase. Furthermore, many relevant training aspects such as hyperparameters were hard-coded and difficult to find. These aspects also made it challenging to expand on the given implementation. The overall provided setup for the implementation failed in our case since the requirements file was not complete and additional bugs were present in the code. Furthermore, the given requirements and batch file could only be run on a Windows machine. Certain adjustments were done on our part in order to run the code on all the necessary local machines.

Finally, we noticed certain deviations between specifications made in the paper and their implementations in the codebase. For instance, the original paper stated that the data was evenly split (the code had a 60-40 split), 25 trials were run for all experiments (the code ran some experiments for only 20 or 10 trials), and interpolation factors of 0.25 and 0.3 were used (the code provided used 0.25 and 0.5). This created confusion in which specifications to apply for the experiments. Because we used the code provided by the authors we decided to give preference to the specifications in the codebase.

Comparing results: An important aspect of our work was to reproduce the paper´s results with the setup provided by the authors. However, they only provided visualizations rather than the raw results, which made it difficult to engage in a precise comparison.

5.3 Communication with original authors

We contacted the first author of the original paper at the beginning of the second week, for instance, to ask for the raw values used for the visualizations in the paper. As of February 3rd, we have not received a response to our email.

References

1. S. Giguere, B. Metevier, B. Castro da Silva, Y. Brun, P. Thomas, and S. Niekum. "Fairness guarantees under demographic shift." In: **International Conference on Learning Representations**. 2022.
2. N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. "A survey on bias and fairness in machine learning." In: **ACM Computing Surveys (CSUR)** 54.6 (2021), pp. 1–35.
3. A. Agarwal, A. Beygelzimer, M. Dudik, J. Langford, and H. Wallach. "A Reductions Approach to Fair Classification." In: **Proceedings of the 35th International Conference on Machine Learning**. Ed. by J. Dy and A. Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, Oct. 2018, pp. 60–69.
4. P. S. Thomas, B. C. da Silva, A. G. Barto, S. Giguere, Y. Brun, and E. Brunskill. "Preventing Undesirable Behavior of Intelligent Machines." In: **Science** 366.6468 (22 November 2019). DOI: 10.1126/science.aag3311, pp. 999–1004.
5. S. Giguere. **Fairness guarentees under demographic shift**. https://github.com/sgiguere/Fairness-Guarantees-under-Demographic-Shift. 2022.
6. R. Kohavi and B. Becker. "Uci machine learning repository: adult data set." In: **Avaliable: https://archive. ics. uci. edu/ml/machine-learning-databases/adult** (1996).
7. B. C. da Silva. **UFRGS Entrance Exam and GPA Data**. 2019.

A Additional results reproducing the original paper

A.1 Adult Dataset

	Known DS				Unknown DS			
	NSF	Acc	FR	Δ Acc	NSF	Acc	FR	Δ Acc
FairConst	n/a	0.784	1.000	-0.004	n/a	0.784	0.250	-0.007
RFLearn	n/a	**0.786**	1.000	-0.000	n/a	**0.788**	1.000	0.001
Fairlearn	n/a	0.750	0.500	-0.002	n/a	0.752	0.500	-0.008
Quasi-SC	0.240	0.765	0.105	0.029	0.000	0.785	**0.000**	0.020
Shifty	0.480	0.756[1]	**0.000**	0.020	0.000	0.780[2]	**0.000**	0.020
SC	0.520	0.756	**0.000**	n/a	0.000	0.781	**0.000**	0.032

[1] significantly worse than best model, $p < 0.001$, $t = 14.256$, $df = 36$
[2] significantly worse than best model, $p < 0.001$, $t = 8.892$, $df = 14$

Table 6. Comparison between the models for the greatest sample size using the Disparate Parity fairness definition and Adult dataset. Shifty differs significantly from the best performing model, and fewer solutions are found. *DS* = Demographic Shift, *NSF* = No Solution Found, *Acc* = Accuracy, *FR* = Failure Rate, Δ *Acc* = Difference in accuracy between the largest and smallest sample size.

	Known DS				Unknown DS			
	NSF	Acc	FR	Δ Acc	NSF	Acc	FR	Δ Acc
FairConst	n/a	0.783	0.040	-0.004	n/a	0.783	**0.000**	-0.007
RFLearn	n/a	**0.786**	0.800	-0.000	n/a	0.787	1.000	-0.001
Fairlearn	n/a	0.783	0.220	0.004	n/a	0.784	0.125	-0.000
Quasi-SC	0.080	0.781	**0.000**	0.046	0.250	0.792	**0.000**	-0.003
Shifty	0.240	0.781[1]	**0.000**	0.044	0.750	0.757[2]	**0.000**	0.025
SC	0.280	0.776	**0.000**	n/a	0.625	**0.825**	**0.000**	n/a

[1] significantly worse than best model, $p = 0.002$, $t = 3.241$, $df = 42$
[2] insufficient number of solutions to perform t-test

Table 7. Comparison between the models for the greatest sample size using the Equalized Odds fairness definition and Adult dataset. When the Demographic Shift (*DS*) is known, Shifty differs significantly from the best performing model.

	Known DS				Unknown DS			
	NSF	Acc	FR	Δ Acc	NSF	Acc	FR	Δ Acc
FairConst	n/a	0.783	**0.000**	-0.005	n/a	0.803	**0.000**	-0.005
RFLearn	n/a	**0.786**	0.440	-0.001	n/a	0.787	1.000	-0.018
Fairlearn	n/a	0.782	0.080	0.001	n/a	**0.852**	0.062	0.016
Quasi-SC	0.160	0.783	**0.000**	0.047	0.000	0.822	**0.000**	0.096
Shifty	0.280	0.783[1]	**0.000**	0.056	1.000	n/a[2]	n/a	n/a
SC	0.440	0.782	**0.000**	0.049	0.500	0.812	**0.000**	n/a

[1] not significantly different from best model, $p = 0.166$, $t = 1.410$, $df = 41$
[2] insufficient number of solutions to perform t-test

Table 8. Comparison between the models for the greatest sample size using the Equal Opportunity fairness definition and Adult dataset. When the *DS* is known, Shifty does not differ significantly from the best performing model.

	Known DS				Unknown DS			
	NSF	Acc	FR	Δ Acc	NSF	Acc	FR	Δ Acc
FairConst	n/a	0.783	0.520	-0.002	n/a	0.792	**0.000**	0.002
RFLearn	n/a	**0.786**	1.000	-0.000	n/a	0.788	1.000	-0.000
Fairlearn	n/a	0.781	0.980	0.003	n/a	0.783	0.750	-0.002
Quasi-SC	0.200	0.779	**0.000**	0.046	0.000	0.789	**0.000**	0.022
Shifty	0.520	0.776[1]	**0.000**	0.046	0.125	0.795[2]	**0.000**	0.031
SC	1.000	n/a	n/a	n/a	0.250	**0.815**	**0.000**	n/a

[1] significantly worse than best model, $p < 0.001$, $t = 5.871$, $df = 35$
[2] not significantly different from best model, $p = 0.201$, $t = 1.360$, $df = 11$

Table 9. Comparison between the models for the greatest sample size using the Predictive Equality fairness definition and Adult dataset. When the *DS* is known, Shifty differs significantly from the best performing model. There is no significant difference when the *DS* is unknown.

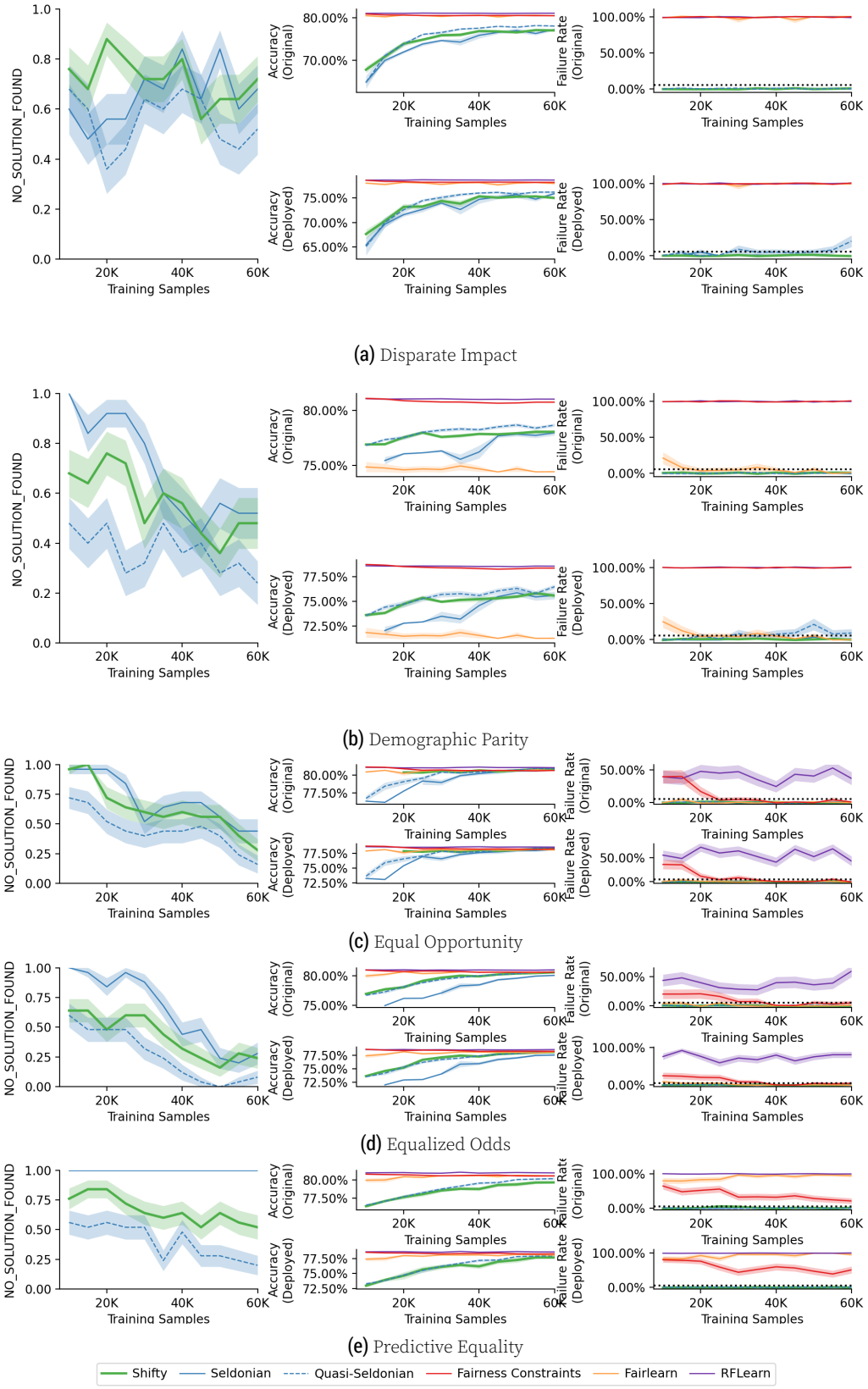

Figure 4. Additional results when enforcing fairness constraints under known demographic shift using the UCI Adult Census dataset.

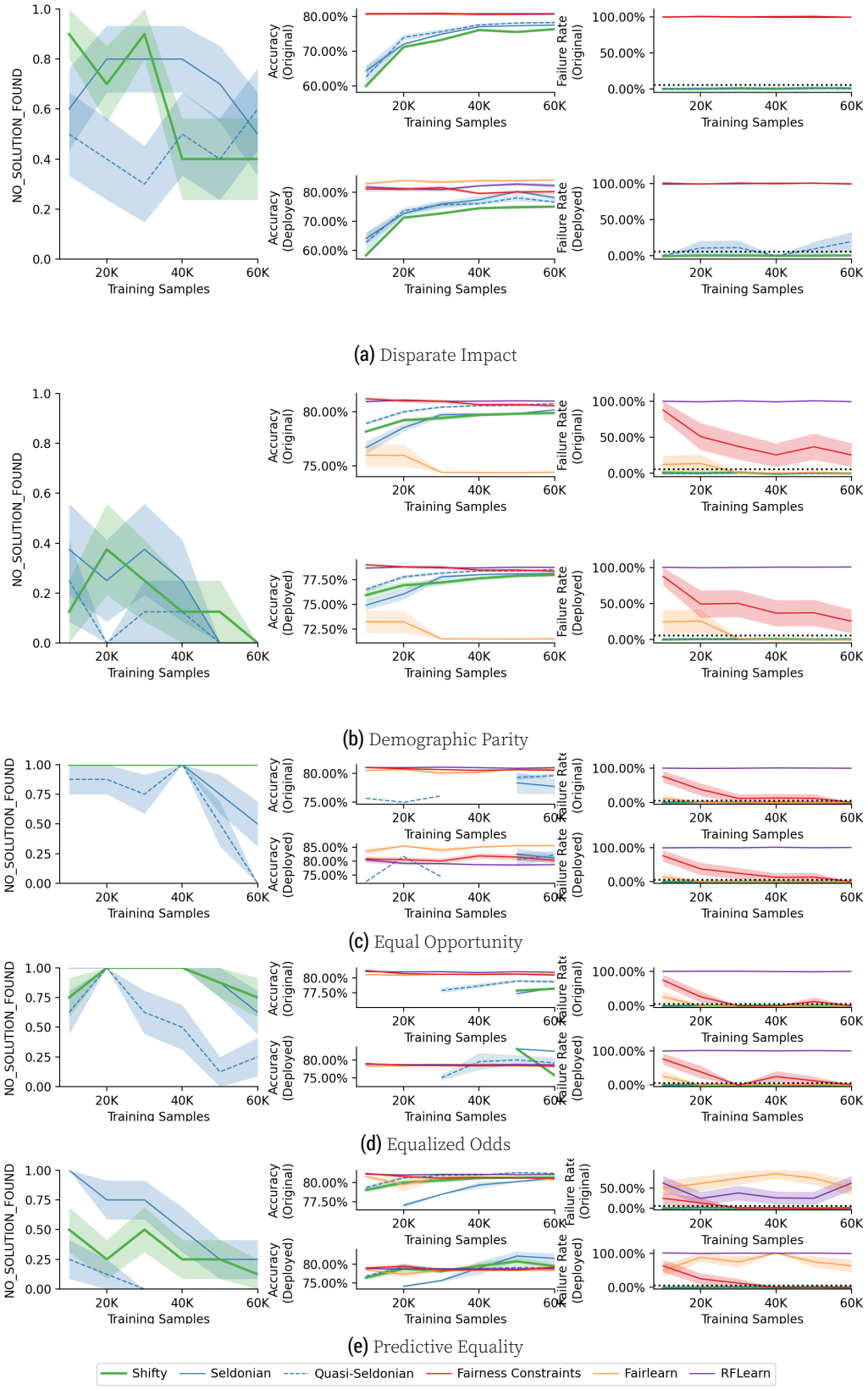

Figure 5. Additional results when enforcing fairness constraints under unknown demographic shift using the UCI Adult Census dataset.

A.2 Brazil Dataset

	Known DS				Unknown DS			
	NSF	Acc	FR	Δ Acc	NSF	Acc	FR	Δ Acc
FairConst	n/a	0.612	1.000	-0.000	n/a	0.612	1.000	0.000
RFLearn	n/a	0.648	**0.000**	0.001	n/a	0.645	**0.000**	0.000
Fairlearn	n/a	0.653	0.400	0.001	n/a	0.649	0.500	0.001
Quasi-SC	0.000	**0.657**	0.160	0.006	0.000	0.655	0.050	0.007
Shifty	0.520	0.643[1]	**0.000**	0.058	0.800	0.628[2]	**0.000**	n/a
SC	0.000	0.656	0.080	0.005	0.000	**0.656**	0.050	0.006

[1] significantly worse than best model, $p < 0.001$, $t = 5.743$, $df = 35$
[2] significantly worse than best model, $p = 0.003$, $t = 3.406$, $df = 22$

Table 10. Comparison between the models for the greatest sample size using the Demographic Parity fairness definition and Brazil dataset. When the *DS* is known, Shifty differs significantly from the best performing model. There is no significant difference when the *DS* is unknown.

	Known DS				Unknown DS			
	NSF	Acc	FR	Δ Acc	NSF	Acc	FR	Δ Acc
FairConst	n/a	0.612	1.000	-0.001	n/a	0.612	1.000	0.001
RFLearn	n/a	0.650	**0.000**	-0.001	n/a	**0.656**	0.900	0.001
Fairlearn	n/a	**0.651**	0.520	-0.000	n/a	0.647	0.475	0.000
Quasi-SC	0.000	0.648	0.080	0.084	0.050	0.650	0.211	0.064
Shifty	0.920	0.570[1]	**0.000**	n/a	1.000	n/a[2]	n/a	n/a
SC	0.240	0.634	**0.000**	0.108	0.350	0.640	**0.077**	0.075

[1] insufficient number of solutions to perform t-test
[2] insufficient number of solutions to perform t-test

Table 11. Comparison between the models for the greatest sample size using the Equalized Odds fairness definition and Brazil dataset.

	Known DS				Unknown DS			
	NSF	Acc	FR	Δ Acc	NSF	Acc	FR	Δ Acc
FairConst	n/a	0.612	1.000	-0.000	n/a	0.614	1.000	0.002
RFLearn	n/a	0.650	0.320	0.002	n/a	0.647	1.000	0.000
Fairlearn	n/a	0.651	0.400	0.000	n/a	0.648	0.550	0.002
Quasi-SC	0.000	0.655	0.040	0.110	0.050	**0.665**	0.895	0.125
Shifty	0.960	0.532[1]	**0.000**	-0.084	0.900	0.493[2]	**0.000**	n/a
SC	0.040	**0.657**	0.042	0.130	0.150	0.663	0.882	0.239

[1] insufficient number of solutions to perform t-test
[2] insufficient number of solutions to perform t-test

Table 12. Comparison between the models for the greatest sample size using the Equal Opportunity fairness definition and Brazil dataset.

	Known DS				Unknown DS			
	NSF	Acc	FR	Δ Acc	NSF	Acc	FR	Δ Acc
FairConst	n/a	0.613	1.000	0.000	n/a	0.612	1.000	0.000
RFLearn	n/a	0.649	0.440	0.001	n/a	0.648	0.150	-0.001
Fairlearn	n/a	0.650	0.900	0.000	n/a	0.648	0.925	-0.000
Quasi-SC	0.000	**0.657**	0.680	0.025	0.000	0.656	0.400	0.014
Shifty	0.880	0.618[1]	**0.000**	-0.009	0.850	0.629[2]	**0.000**	-0.011
SC	0.040	0.656	0.375	0.129	0.200	**0.658**	**0.000**	0.092

[1] significantly worse than best model, $p < 0.001$, $t = 6.943$, $df = 26$
[2] significantly worse than best model, $p = 0.008$, $t = 2.981$, $df = 17$

Table 13. Comparison between the models for the greatest sample size using the Predictive Equality fairness definition and Brazil dataset. Shifty performs significantly worse than the best model for both the known and unknown demographic shift.

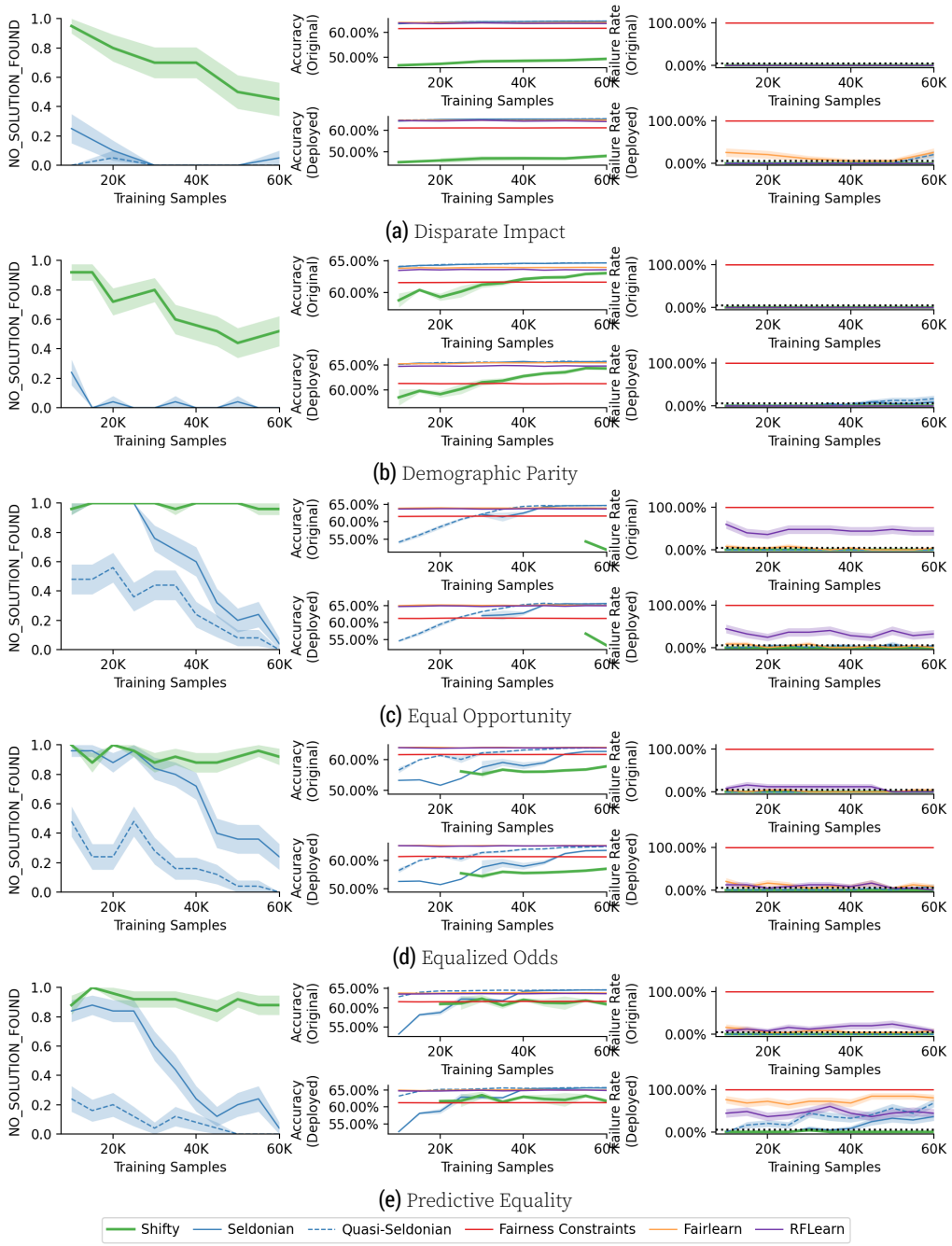

Figure 6. Additional results when enforcing fairness constraints under known demographic shift using the UFRGS Entrance Exam and GPA dataset.

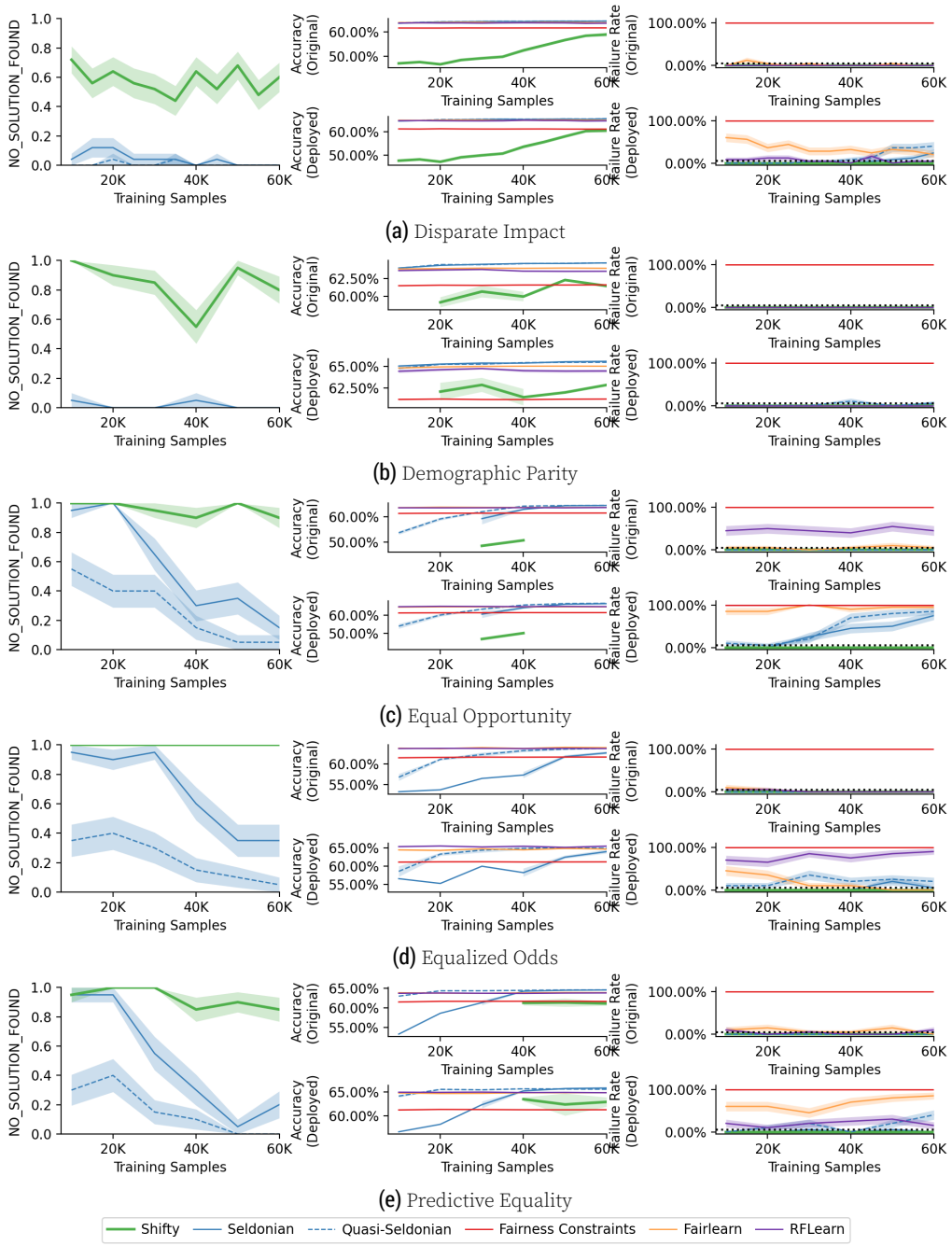

Figure 7. Additional results when enforcing fairness constraints under unknown demographic shift using the UFRGS Entrance Exam and GPA dataset.

A.3 Additional results beyond original paper

	Known DS				Unknown DS			
	NSF	Acc	FR	Δ Acc	NSF	Acc	FR	Δ Acc
CMA-ES	0.600	**0.606**	**0.000**	0.130	0.450	**0.480**	**0.000**	0.030
BFGS	0.440	0.596	**0.000**	0.120	0.050	0.469	**0.000**	0.019

Table 14. Comparison between CMA-ES and BFGS optimization methods for the *Shifty* implementation using the Brazil dataset and the Disparate Impact fairness definition.

	Known DS				Unknown DS			
	NSF	Acc	FR	Δ Acc	NSF	Acc	FR	Δ Acc
CMA-ES	0.480	**0.756**	**0.000**	0.020	0.000	**0.780**	**0.000**	0.020
BFGS	0.240	0.731	**0.000**	0.014	0.100	0.755	**0.000**	0.016

Table 15. Comparison between CMA-ES and BFGS optimization methods for the *Shifty* implementation using the UCI Adult Census dataset and the Demographic Parity fairness definition.

	Known DS				Unknown DS			
	NSF	Acc	FR	Δ Acc	NSF	Acc	FR	Δ Acc
CMA-ES	0.520	**0.643**	**0.000**	0.058	0.800	**0.628**	**0.000**	n/a
BFGS	0.480	0.638	**0.000**	0.079	0.600	0.613	**0.000**	0.034

Table 16. Comparison between CMA-ES and BFGS optimization methods for the *Shifty* implementation using the Brazil dataset and the Demographic Parity fairness definition.