Fairness without Demographics through Adversarially Reweighted Learning: Reproduction and Extension Report

Tharak Koneni KTHARAK@USF.EDU
Department of Computer Science
University of South Florida
Tampa, FL, USA

Tarun Reddy Boreddy BOREDDY@USF.EDU
Department of Computer Science
University of South Florida
Tampa, FL, USA

Sai Balaji Reddy Karumuri SAIBALAJIREDDY@USF.EDU
Department of Computer Science
University of South Florida
Tampa, FL, USA

Abstract

Fairness in machine learning models is a critical concern, especially in high-stakes decision-making domains. In this project, we replicate the "Fairness without Demographics through Adversarially Reweighted Learning" (ARL) framework, originally presented at NeurIPS 2020. We successfully reproduce the ARL method on the datasets given in the paper i.e., COMPAS, UCI-Adult and LSAC and further extend the evaluation to two new datasets: the Police Killings and Facebook datasets. We highlight the challenges encountered during reproduction and new experimentation, analyze subgroup-specific fairness metrics, and reflect on reproducibility gaps. Our findings emphasize the practical difficulties in achieving fairness without demographic labels and offer insights for future research.

1 Introduction

As machine learning (ML) systems are increasingly deployed in sensitive and high-stakes domains—such as criminal justice, healthcare, finance, and hiring—there is growing concern about algorithmic bias and fairness. Ensuring that ML models do not systematically disadvantage individuals based on protected attributes (like race, gender, or socio-economic background) is both a technical and ethical imperative. Traditional group fairness approaches often rely on access to explicit demographic labels to enforce parity constraints or evaluate disparate impact. However, in many real-world scenarios, such demographic information is either unavailable due to privacy regulations or difficult to collect reliably.

To address this gap, the paper "Fairness Without Demographics Through Adversarially Reweighted Learning" Lahoti et al. (2020), proposes a method that enforces fairness without requiring explicit demographic labels at training time. The core idea of their approach, termed Adversarially Reweighted Learning (ARL), is to learn instance weights using an adversarial model that maximizes subgroup AUC gaps, thereby indirectly encouraging the main classifier to perform more equitably across unknown subgroups.

This project focuses on reproducing the results from that paper, with a primary goal on the ARL algorithm, since it is the central contribution and innovation of the work. Specifically, we reimplemented the ARL approach as described in the original paper, and evaluated its fairness and accuracy performance on the original datasets i.e., COMPAS, UCI Adult and LSAC Law School Admissions datasets which were studied in the original paper. In addition, we extended the experimentation to two new real-world datasets i.e.,Police Killings, and Facebook Data. This allowed us to test the generalizability and robustness of ARL across diverse contexts with different types of subgroup structures and outcome imbalances.

Our primary objective is to assess whether the ARL method retains its fairness advantages in realistic and potentially noisy settings. We report performance in terms of AUC average, minority group AUC, macro-averaged AUC, and AUC minimum, aligning with the original evaluation framework. The analysis also includes implementation challenges, results reproduction, and key takeaways regarding the broader applicability of fairness without demographics.

2 Related Work

The problem of ensuring fairness in machine learning has been widely studied, giving rise to a broad taxonomy of fairness definitions and algorithmic interventions. Much of the early work focuses on group fairness metrics such as demographic parity, equalized odds which evaluate model performance across predefined demographic groups. Approaches in this category often enforce fairness constraints during training (Hardt et al. (2016); Zafar et al. (2017)). However, these techniques critically rely on access to protected group attributes, which are not always available or usable in real-world datasets due to privacy concerns or data collection limitations.

More recently, Distributionally Robust Optimization (DRO)Hashimoto et al. (2018) approaches have been introduced to improve group fairness by minimizing the worst-case loss over predefined demographic groups. While these methods can effectively reduce disparities in group performance, they inherently rely on knowing the group identities in order to define and compute group-wise loss terms. This dependence on demographic information restricts their applicability, especially in contexts where group labels are unavailable due to privacy, regulatory, or operational constraints.

To address these limitations, Lahoti et al. (2020) propose Adversarially Reweighted Learning (ARL), a novel group-agnostic technique that does not require access to group labels at any stage. ARL uses an adversarial framework to reweight training samples based on model loss, implicitly prioritizing underperforming subpopulations. This is achieved by training a weight generator adversary that learns to assign importance weights to data points to maximize model loss, while the main model attempts to minimize the weighted loss. This adversarial reweighting simulates the effect of DRO in a group-agnostic way, thereby improving fairness without relying on demographic attributes.

Their work builds on the idea that examples with higher loss during training are often indicative of underrepresented or harder to predict subgroups, and leverages this to improve model performance across such latent groups. Through comprehensive evaluations on multiple datasets including COMPAS, UCI Adult, and LSAC, ARL is shown to match or outperform group-aware baselines in minimizing disparities across subgroups, while requiring no demographic information during training or testing.

3 Experimental Design

We used the ARL framework proposed by the authors, which was implemented through TensorFlow. The two main components of this setup are learner and Adversary. The learner model which is the primary classifier is a fully connected neural network consisting of two hidden layers with 64 and 32 units respectively, with ReLu activation function. The adversary model is a linear classifier. Adversary assigns example weights based on computationally identifiable error regions.

The ARL framework is inspired by the Rawlsian min-max principle, which originates from political philosophy and is commonly applied in machine learning fairness to promote equity across groups. The core idea is to optimize performance for the worst-off group, to maximize the minimum utility across all subgroups.The authors formulated ARL as the MinMax optimization problem between the learner and adversary. The learner tries to learn best parameter θ that minimizes expected loss.

The adversary maximizes this loss by focusing weight on samples where the classifier performs poorly. These weights are dynamically updated by the adversary using a sigmoid-activated scoring function, which is normalized per training batch. The adversary assigns the weight to computationally identifiable error zone where the learner makes significant errors, thus improving the learners performance. Moreover, the adversary does not use demographic information, its goal is to uncover computationally identifiable subgroups through patterns in the data that correlate with prediction errors. This indirect strategy is what enables ARL to enforce fairness without access to sensitive group labels.

For all the datasets, Authors performed 5-fold cross validation to select hyper parameters such as learning rate, batch size and dropout. The loss function used is standard binary cross-entropy. Evaluation metrics include accuracy and AUC but the authors chose AUC (area under the ROC curve) as their primary metric because it is robust to class imbalance, considers both false positive rate (FPR) and false negative rate (FNR), and is not dependent on a specific threshold. AUC(min) (minimum AUC over all subgroups) and AUC(minority) (AUC for the smallest subgroup) are the other two important metrics used to evaluate ARL. Author compared AUC scores of ARL with previous work such as DRO, IPW, and Min-Diff for all datasets.

4 Datasets

We evaluated ARL on three publicly available datasets from the original paper and extended the evaluation to two additional datasets. Each dataset presents a unique challenge in terms of feature diversity, label imbalance, and subgroup structure.

4.1 Original Datasets

Adult(UCI): The Adult dataset, sourced from the UCI Machine Learning Repository, contains census income information used for binary classification. The objective is to predict whether an individual earns more than $50K annually. After preprocessing, the dataset includes 40,701 samples and 15 features. The target variable is income (>50K or <=50K). We used an 80/20 train-test split. Protected attributes include race and sex, leading to subgroups such as White-Male, White-Female, Black-Male, and Black-Female.(UCI Machine Learning Repository (1996a,b)).

LSAC: The LSAC dataset comprises law school applicant records, capturing demographic and academic features. The task is to predict whether a student passed the bar exam. It contains 27,479 samples and 12 features. The target variable is binary (`passed` or `failed`). Similar to Adult, we used an 80/20 train-test split. Protected groups are defined by combinations of **race** and **sex**. We also analyzed the LSAC dataset (Ofer (2021)).

COMPAS: The COMPAS dataset includes criminal history, demographic, and recidivism risk scores. The prediction task is whether a defendant re-offended within two years. The dataset contains 7,215 samples and 11 features after preprocessing. The target is binary (`recidivated` or `did not recidivate`). An 80/20 split was used for training and testing. Protected groups are derived from **race** and **sex**, as used in the original ARL evaluation. ProPublica (2016).

4.2 New datasets

Police Killings (US): This dataset captures information about individuals killed by police in the United States, with contextual features such as age, armed status, race, and location. After preprocessing, we retained 4,000 samples and selected a binary target variable representing whether a lethal force incident occurred under specific high-risk circumstances. The dataset includes both categorical and numerical features, and we used an 80/20 train-test split. Since protected attributes were not explicitly used, subgroup fairness was evaluated based on computationally identified subpopulations. We used the police killings dataset (Wullum (2022)).

Facebook: This dataset comprises detailed information about posts made on the Facebook page of a renowned cosmetic brand during 2014. It includes various attributes such as post type, category, and engagement metrics such as likes, comments, and shares. The primary objective is to analyze factors that influence post engagement. After preprocessing, we retained 100,000 samples. An 80/20 train-test split was utilized. Protected attributes were not explicitly available; thus, subgroup fairness was assessed based on computationally identified subpopulations. (Batra (2022)).

5 What Was Easy and What Was Hard

5.1 Easy:

Several aspects of the reproduction process were relatively straightforward. The model architecture itself is simple: the learner is a 2-layer feed-forward neural network with ReLU activation, and the adversary is a linear model, which made it easy to understand and implement. Publicly available datasets such as Adult, LSAC, and COMPAS were easy to access (except for LSAC, which required sourcing from Kaggle) and integrate into the training pipeline. The use of standard evaluation metrics like AUC further facilitated benchmarking. Moreover, the provided computational graph clearly illustrated the interaction between the learner and the adversary, simplifying the conceptual understanding of the adversarial reweighting mechanism. Running the original ARL codebase was smooth due to its good structure, and preprocessing the data was manageable after aligning features to the expected format.

5.2 Hard:

The process was not without challenges. Setting up the environment required using Python 3.6.8, which was not mentioned in the original codebase. Later versions of Python led to compatibility issues, particularly with TensorFlow dependencies, and identifying the correct version involved trial and error. The instructions in the README file were insufficiently detailed regarding how to execute training and testing, which forced us to reverse-engineer usage from the source code. Dataset variability introduced further complexity especially in the COMPAS dataset, where weak protected-group signals made it harder for ARL to identify fairness-relevant regions. Additionally, the original LSAC dataset link cited in the paper was expired, requiring us to manually verify and obtain it from alternative sources like Kaggle.

To produce the results for the new datasets, we had to create new input pipelines: `facebook_input.py` and `police_input.py`, by referring the original `uciadult_input.py` provided in the ARL codebase. These custom scripts defined how data was read, preprocessed, and formatted for the ARL framework. We maintained consistency with the original format by incorporating nearly all elements from the reference code, including standardized parsing of numerical and categorical features, generation of the `mean_std.json` and `vocabulary.json` files, and alignment of target labels and batch construction.This effort required detailed understanding of the original dataset interface, particularly to support consistent group-agnostic evaluation and IPS weight handling for the Facebook and Police Killings datasets. Implementing these input functions from scratch while preserving compatibility with the adversarial training pipeline was essential for ensuring valid and comparable fairness evaluations. Finally, understanding ARL's minimax adversarial loss formulation required in-depth study.

6 Reproduced Results on Original Datasets

As mentioned earlier, Author compared AUC scores of ARL to existing related work like DRO,IPW,Min-Diff. ARL outperforms IPW in all the AUC scores across all three datasets. ARL outperform DRO on Adult and LSAC datset, however due to Label noise and weak Correlation between protected groups and feature space DRO does better on COMPAS dataset as it has access to protected group, which ARL doesn't use.
These are results we obtained upon training and testing the datasets using ARL.

Table 1: Comparison of AUC Metrics Across Datasets (Paper vs. Our Results)

Dataset	Method	AUC avg	AUC macro-avg	AUC min	AUC minority
COMPAS	Paper	0.743	0.727	0.658	0.785
	Our Result	0.745	0.727	0.663	0.665
LSAC	Paper	0.823	0.820	0.798	0.832
	Our Result	0.794	0.759	0.740	0.793
UCI Adult	Paper	0.907	0.915	0.881	0.942
	Our Result	0.903	0.889	0.863	0.924

We reproduced the results on the three original datasets used in the ARL paper: Adult, LSAC, and COMPAS. Each dataset exhibited unique characteristics in terms of group separability and fairness response to the ARL framework.

On the Adult dataset, our reproduced results were very close to those reported in the paper, with AUC(min) reaching 0.863 and AUC(minority) 0.924. This high performance is also visible in the original findings and can be attributed to the well-defined protected subgroups (race and sex), as well as the large dataset size and clean feature-label relationship. Among all datasets, Adult yielded the best fairness metrics, likely due to the clear and linearly separable subgroup structures. Also the paper showed that there is stronger correlation between protected attributes and other input features which is very helpful for ARL to efficitively isolate and upweight underperforming subgroups

For the LSAC dataset, we achieved AUC(min) of 0.740 and AUC(minority) of 0.793. These values are generally aligned with those in the original paper, though slight differences were observed. One potential reason for the variation is that the original LSAC dataset URL is no longer active, and we had to rely on a third-party copy hosted on Kaggle. While we verified the structure and key statistics of the downloaded dataset, minor differences in formatting or preprocessing may have influenced the results.Still, the adversarial weighting strategy was able to detect group disparities based on race and sex and improved fairness across subgroups. Compared to Adult, LSAC showed slightly lower fairness gains, possibly due to fewer training examples and more nuanced subgroup dynamics. The model still demonstrated improved fairness, confirming ARL's effectiveness on this dataset.

The COMPAS dataset proved to be the most challenging. Our reproduced results showed AUC(min) of 0.663 and AUC(minority) of 0.665, which are comparable to the original findings but did not show as dramatic a fairness gain as seen with Adult. This is likely due to two key factors: First, the protected group information (race and sex) is not easily inferable from the feature set, as evidenced by low group prediction accuracy using linear models; and second, the COMPAS labels themselves are noisy and potentially biased, making it difficult for the adversary to distinguish genuine error-prone regions from mislabeled data. These challenges make it difficult for ARL to effectively upweight disadvantaged subgroups. As a result, among all datasets, COMPAS showed the weakest fairness improvements despite ARL's design for group-agnostic fairness.

7 Results on New Datasets

Table 2: Comparison of AUC Metrics Across Datasets (Paper vs. Our Results)

Dataset	**AUC avg**	**AUC macro-avg**	**AUC min**	**AUC minority**
Police	0.490	0.894	0.000	0.000
Facebook	0.935	0.929	0.915	0.875

To assess the generalizability of the ARL approach beyond the datasets used in the original study, we extended our evaluation to two additional real-world datasets: the Police Killings (US) dataset and the Facebook Comment Toxicity dataset.

On the Police Killings dataset, ARL achieved a relatively high overall accuracy of 0.875 but yielded a very low AUC average (0.49) and zero values for AUC(min) and AUC(minority), indicating poor subgroup fairness. One contributing factor to this poor performance is the extreme imbalance in subgroup representation. Specifically, there are only around 100 samples associated with the female group. Such a small sample size limits the adversary's ability to detect and prioritize fairness improvements for underrepresented subpopulations, which diminishes ARL's subgroup-targeting effectiveness. This imbalance may be a key reason for the weak fairness performance observed, as the adversarial reweighting mechanism relies on sufficient representation across subgroups to generalize effectively. This suggests that while the model performed well on average, it failed to generalize to computationally-identifiable worst-case subgroups.

In contrast, on the Facebook dataset, ARL demonstrated strong performance both overall and across subgroups. The model achieved an accuracy of 0.887, with AUC(avg) of 0.935, AUC(macro-avg) of 0.929, and high fairness metrics: AUC(min) at 0.915 and AUC(minority) at 0.875. While these results indicate effective subgroup handling, we note that the dataset includes protected attribute labels such as gender, religion, and race, which allowed for explicit subgroup fairness evaluation. This structure enabled the adversary to detect underperforming subgroups more effectively. However, the large sample size, 100,000 records provided adequate representation across most subgroups, enabling the adversary to learn more stable and effective reweighting strategies during training.Overall, Facebook served as a strong validation case for ARL, demonstrating that when group membership can be inferred and when group-related structure is computationally identifiable ARL can substantially improve fairness without sacrificing accuracy.

8 Key Takeaways and Conclusion

8.1 Key Takeaways

Adversarially Reweighted Learning (ARL) enhances fairness without the need for protected demographic attributes by dynamically reweighting training instances that lie in high-error regions of the input space. A key advantage of ARL is its ability to function without demographic supervision during both training and inference, which makes it particularly suitable for real-world contexts with strict privacy regulations or missing group labels.

Rather than identifying explicitly disadvantaged groups ARL prioritizes regions of the data distribution where the model underperforms, thereby improving the performance for hidden or unobserved minority subgroups. Across our experiments, ARL consistently outperformed group-aware baselines like DRO, IPW, and Min-Diff in terms of fairness metrics such as AUC(minority) and AUC min, especially in datasets where group disparities are indirectly encoded in the feature-label relationships (e.g., UCI Adult, Facebook).

However, we observed that ARL is sensitive to label noise and model miscalibration. On noisier datasets like COMPAS, which may contain weak or indirect correlations between features and outcomes, ARL's subgroup fairness performance showed marginal degradation. These observations align with the original paper's claim that ARL's strength lies in its ability to discover and exploit computationally identifiable error regions.

From an engineering standpoint, reproducing ARL presented several technical challenges. The original codebase was based on deprecated TensorFlow 1.x components, requiring environment setup with Python 3.6 and TensorFlow 1.13.2, along with extensive code fixes and dependency management. Despite these hurdles, ARL was successfully reimplemented across five datasets (COMPAS, LSAC, UCI Adult, Facebook, Police Killings), without using protected group labels in training. The method showed strong generalizability, achieving reliable subgroup fairness across both synthetic benchmarks and real-world domains.

8.2 Conclusion

This project aimed to reproduce and extend the fairness methodology introduced in Fairness without Demographics through Adversarially Reweighted Learning (Lahoti et al. (2020)). The central focus was on evaluating the performance and robustness of ARL, a fairness framework designed to optimize subgroup equity without needing access to demographic attributes.

Our experiments demonstrate that ARL achieves significant improvements in fairness metrics such as AUC min, AUC minority, and macro AUC across a range of datasets. These results held true not only in standard benchmarks like COMPAS, LSAC, and UCI Adult, but also in complex, real-world datasets like Facebook and Police Killings, where protected group information is not explicitly available. The performance gap between paper results and our reimplementation was minimal in most cases, suggesting that ARL is a reproducible and robust approach.

In essence, ARL offers a practical and scalable solution to bias mitigation in scenarios where demographic data is unavailable, incomplete, or ethically restricted. By shifting the fairness paradigm from group-based correction to error-based correction, ARL opens a new direction for equitable AI development under privacy constraints. This work confirms ARL's viability and highlights its promise for future research in fair and accountable machine learning.

Appendix A.

The Adversarially Reweighted Learning (ARL) objective can also be expressed in its discrete form as follows:

$$J(\theta, \phi) = \min_{\theta} \max_{\phi} \sum_{i=1}^{n} \lambda_{\phi}(x_i, y_i) \cdot \ell(h_{\theta}(x_i), y_i)$$

Here:

- θ are the parameters of the learner h_{θ},
- ϕ are the parameters of the adversary that computes the weights $\lambda_{\phi}(x_i, y_i)$,
- ℓ is the loss function (e.g., binary cross-entropy),
- n is the number of training examples.

This formulation captures the learner-adversary dynamics, where the adversary highlights error-prone regions through λ_{ϕ}, and the learner adapts to improve performance on these reweighted samples.

Appendix B.

Although not included in the main discussion, we initially evaluated ARL on the UCI dataset titled *Estimation of Obesity Levels Based on Eating Habits and Physical Condition* (link). The dataset contains 2,111 instances with 17 features (16 inputs + 1 target label).Since the police dataset (Wullum (2022)) was a small dataset too, to test the ARL performance, we went with facebook dataset which is much larger (Batra (2022)). After preprocessing and applying ARL, we observed the following performance:

- **AUC(avg)**: 0.885
- **AUC(min)**: 0.845

References

Sheena Batra. Facebook data. `https://www.kaggle.com/datasets/sheenabatra/facebook-data`, 2022. Accessed via Kaggle API; login required. Accessed: April 2025.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. *CoRR*, abs/1610.02413, 2016. URL `http://arxiv.org/abs/1610.02413`.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1929–1938. PMLR, 10–15 Jul 2018. URL `https://proceedings.mlr.press/v80/hashimoto18a.html`.

Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed H. Chi. Fairness without demographics through adversarially reweighted learning. *CoRR*, abs/2006.13114, 2020. URL `https://arxiv.org/abs/2006.13114`.

Dan Ofer. Law school admissions and bar passage dataset. `https://www.kaggle.com/datasets/danofer/law-school-admissions-bar-passage`, 2021. Accessed via Kaggle; requires API key or login to download. Accessed: April 2025.

ProPublica. Compas scores - two years. `https://raw.githubusercontent.com/propublica/compas-analysis/refs/heads/master/compas-scores-two-years.csv`, 2016. Raw data file.

UCI Machine Learning Repository. Adult data set - training file. `https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data`, 1996a. Accessed: April 2025.

UCI Machine Learning Repository. Adult data set - test file. `https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.test`, 1996b. Accessed: April 2025.

Kwame Wullum. Fatal police shootings in the us. `https://www.kaggle.com/datasets/kwullum/fatal-police-shootings-in-the-us`, 2022. Accessed via Kaggle API; login required. Accessed: April 2025.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 1171–1180, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee. ISBN 9781450349130. doi: 10.1145/3038912.3052660. URL `https://doi.org/10.1145/3038912.3052660`.