A Benchmark for Client-level Fairness in Federated Learning

XENIA HEILMANN*, Johannes Gutenberg University, Germany
LUCA CORBUCCI*, University of Pisa, Italy
MATTIA CERRATO, Johannes Gutenberg University, Germany

Federated Learning (FL) enables collaborative model training while preserving participating clients' local data privacy. However, the diverse data distributions across different clients can exacerbate fairness issues, as biases in client data may propagate across the federation. Although various approaches have been proposed to enhance fairness in FL, they typically focus on mitigating the bias of a single binary-sensitive attribute. This narrow focus often overlooks the complexity introduced by clients with conflicting or diverse fairness objectives. Such clients may contribute to the federation without experiencing any improvement in their model's performance or fairness regarding their specific sensitive attributes. To evaluate fair FL methods for global and individual client fairness in a reproducible and reliable manner, the need for standardized datasets becomes apparent. In this paper, we propose a preliminary framework to create benchmarking datasets that allow researchers and developers to evaluate fair FL methods. These benchmarking datasets include various heterogeneous client settings with regard to data bias of sensitive attributes for assessing fairness at the global and individual level. Additionally, we provide information on how to evaluate results obtained with these benchmarking datasets.

Keywords: Benchmark, Fair Federated Learning, Heterogeneity, Federated Learning

Reference Format:
Xenia Heilmann, Luca Corbucci, and Mattia Cerrato. 2025. A Benchmark for Client-level Fairness in Federated Learning. In *Proceedings of Fourth European Workshop on Algorithmic Fairness (EWAF'25).* Proceedings of Machine Learning Research, 5 pages.

1 Introduction

The widespread adoption of Machine Learning (ML) has driven the introduction of numerous AI regulations [16, 11, 3, 7, 13] and an increased focus on Responsible AI. Recent efforts have prioritized explainability [4], user privacy [15], and fairness in ML models [5].

Federated Learning (FL) [14] enhances privacy by enabling collaborative ML training without sharing raw data. Addressing fairness in FL, however, proved complex due to client heterogeneity as clients may have region-specific data biases, differing fairness definitions, and participation motivations. Sensitive attributes like gender, race, or age, often linked to societal biases, require careful consideration to ensure fairness. Yet, existing fair FL approaches

*Both authors contributed equally to this research.

Authors' Contact Information: Xenia Heilmann, Johannes Gutenberg University, Mainz, Germany, xenia.heilmann@uni-mainz.de; Luca Corbucci, University of Pisa, Pisa, Italy, luca.corbucci@phd.unipi.it; Mattia Cerrato, Johannes Gutenberg University, Mainz, Germany, mcerrato@uni-mainz.de.

EWAF'25, June 30–July 02, 2025, Eindhoven, NL

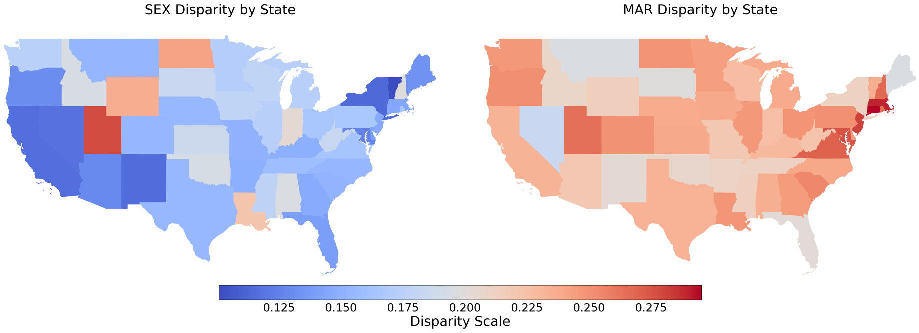

Fig. 1. The fairness measured using Demographic Disparity on the ACS Income dataset. If we consider SEX and MAR as two separate sensitive attributes, we can observe that some states are more biased with respect to one attribute than to the other.

typically assume a single binary sensitive attribute, which may not align with all clients' fairness concerns. While some clients may still benefit from the FL's fairness constraints, others may experience increased unfairness for unaddressed attributes. Bias propagation across a federation [6] can even result in less fair models for some clients than locally trained models, which can lead to less motivation of clients to contribute to federations if no benefit in participation is apparent. Ideally, objectives, including accuracy and fairness metrics, should be clearly defined before training to counteract this.

When FL participation is mandatory due to political or socioeconomic factors, different client settings in terms of e.g., bias distribution, fairness objectives, and data collection are the rule. Yet, most fair FL methods prioritize the mitigation of a global fairness metric computed on the server side. However, they do not consider client-level fairness, not taking into account whether it is worthwhile for clients to be part of the federated training process. To evaluate fair FL methods, it is advantageous to extend the current focus on the global model's fairness to also include an individual fairness evaluation. Specifically, we propose that fair FL methods should be evaluated in settings where

(1) Clients have data biased towards the same sensitive attributes and the same values of these attributes.
(2) Clients have data biased toward different values of the same sensitive attribute.
(3) Clients have data biased toward different sensitive attributes than those addressed by the federation.

Settings (2) and (3) are illustrated in Figure 1 by showing how Demographic Disparity (detailed in Section 2) for Gender (SEX) and Marital Status (MAR) varies across states using the ACS Income dataset [8].

In this paper, we highlight the need for a set of benchmarking datasets that can be used to evaluate the effectiveness of fair FL methods not only at the server but also at the client level. Moreover, we propose a preliminary pipeline that can generate new benchmarking datasets based on practitioners' needs and preferences. We believe that the existence of such a methodology can increase the reproducibility of fair FL research as well as extend its applicability. Lastly, this can also be a solution for the lack of standardized datasets that can be used by all researchers proposing methods that have to deal with the evaluation or mitigation of fairness in FL.

2 Related Work

Evaluating a client's benefit from participating in FL training is essential for understanding model performance on individual client data beyond global test set metrics [18, 9]. While accuracy and F1-score are commonly used to evaluate the quality of a model and the worthiness of taking part in the FL process, no prior work has examined FL participation benefits from a fairness perspective. Several relevant works have proved that FL is highly sensitive to bias propagation. This finding was first highlighted in a preliminary study [10] and then explored in depth in [6]. For an overview of fair FL methods, we refer the reader to the survey by [17]. Here, [17] lists in Table 4 the datasets used in works on group fairness in FL. This Table lists 25 different datasets, with Adult [2] being the most commonly applied one. Furthermore, the authors identify the predominant use of traditional tabular Machine Learning datasets for training and evaluation of fair FL methods. They stress that to ensure that fairness mitigation solutions are widely applicable across diverse practical contexts, it is necessary to provide datasets beyond the traditional ones.

3 Framework Proposal

To fairly evaluate new fair FL methods and enable meaningful comparisons, it is essential to go beyond global fairness metrics, typically measured on the server for a single sensitive attribute, examining how these methods impact individual clients' fairness. We propose to implement a framework to develop benchmarking datasets and evaluation guidance to address this gap and analyze fairness interventions at both levels: global and local. In Figure 2, we provide an overview of our proposal. Here, we envision as a first step to generate multiple benchmarking datasets from a predefined set of configurable parameters. As a second step, we create and evaluate locally trained fair models for each client so that the results obtained with these can be later used for evaluation of the method's local fairness improvements for individual clients.

3.1 Benchmarking Datasets

At the core of our proposal is a data generation pipeline built on real-world datasets. By leveraging the ACS dataset and the Folktables library [8], we propose to allow practitioners to generate multiple benchmarking datasets tailored for fairness evaluations in FL based on their preferences and use cases, with a particular focus on evaluation and baselines at the local level. To generate these datasets, our framework would provide several parameters with various attributes to choose from. In particular, we would allow choosing **(I)** the data sampling strategy, ranging from uniformly random, stratified to non-iid, w.r.t. to $\mathbb{P}(Y=y), \mathbb{P}(Y=y|Z=z), \mathbb{P}(Z=z|Y=y), \mathbb{P}(Z=z)$ **(II)** the number of clients that will be involved in the simulation and will need training data, **(III)** the fairness metric used to evaluate the model unfairness, **(IV)** the sensitive attributes against which we want to measure the unfairness, **(V)** the level of heterogeneity of the dataset distribution **(VI)** the base dataset to be used in the generation of client-level data, chosen among the ones provided in the ACS dataset **(VII)** the FL setting that we want to consider (Cross Silo or Cross Device [12]), **(VIII)** the unfairness level between different clients and their sensitive attributes.

These datasets enable researchers to systematically study fairness across diverse FL scenarios, incorporating multiple sensitive attributes at both global (server-side) and local (client-side) levels. Reproducibility will be a priority of our framework; in particular, we aim to achieve a consistently replicable and comparable fairness evaluation across studies, which can be realized by sharing the chosen parameters to generate the benchmarking

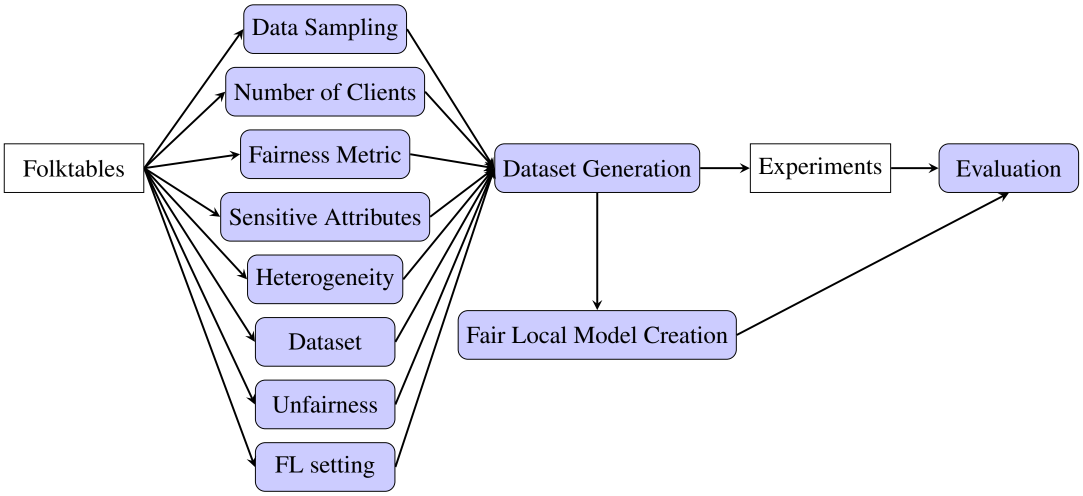

Fig. 2. Pipeline for creating datasets to evaluate fair FL methodologies as well as provide evaluation specifically for individual clients' fairness benefits. Blue is what we integrate into our pipeline.

datasets. This also addresses concerns recently raised regarding the availability and reliability of FL research datasets [1].

3.2 Evaluation

We also want to highlight the importance of a standardized evaluation practice. Here, we propose to evaluate global fairness at the server level and client-specific fairness disparities by comparing fairness (across various sensitive attributes) and accuracy of local fair models (trained on each client) vs. the fair FL model, models trained on centralized data vs. the fair FL model and FL model vs. the fair FL model.

4 Conclusion and Future Work

With this proposal, we aim for a unified framework to provide benchmarking datasets for analyzing the global and local fairness of individual clients in fair FL methods. Our goal is to allow researchers and practitioners to easily test their fair FL solutions on datasets generated specifically for FL settings instead of relying on traditional datasets. The generation of the datasets is based on the well-known ACS Dataset and the Folktables library, which would allow us to work on data collected in the USA that already contains biases, which can also be exacerbated. The existence of such a pipeline would increase the reproducibility of the fair FL methods, allowing researchers to easily compare them on a common benchmark. In future work, we plan to implement all components of the proposed methodology, extending it to generate FL datasets with and without fairness constraints, and evaluate our tool on a subgroup of existing fair FL methods. Another consideration would be to extend the proposed pipeline to synthetic biased data generation [1], which would give additional opportunities for experimentation.

[1] Federated Datasets in Research https://flower.ai/blog/2024-12-02-federated-datasets-in-research/

Acknowledgments

The research in this paper was supported by the “TOPML: Trading Off Non-Functional Properties of Machine Learning” project funded by Carl Zeiss Foundation, grant number P2021-02-014. Also, funded by the European Union; Grant Agreement no. 101120763 - TANGO. Views and opinions expressed are, however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them.

References

[1] J. Baumann et al. “Bias on Demand: A Modelling Framework That Generates Synthetic Data With Bias”. In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’23. Chicago, IL, USA: Association for Computing Machinery, 2023, 1002–1013. ISBN: 9798400701924. DOI: 10.1145/3593013.3594058. URL: https://doi.org/10.1145/3593013.3594058.

[2] B. Becker et al. *Adult*. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20. 1996.

[3] J. R. Biden. “Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence”. In: (2023).

[4] F. Bodria et al. “Benchmarking and survey of explanation methods for black box models”. In: *Data Mining and Knowledge Discovery* 37.5 (2023), pp. 1719–1778.

[5] S. Caton et al. “Fairness in Machine Learning: A Survey”. In: *ACM Comput. Surv.* (). URL: https://doi.org/10.1145/3616865.

[6] H. Chang et al. “Bias propagation in federated learning”. In: *ArXiv preprint* abs/2309.02160 (2023). URL: https://arxiv.org/abs/2309.02160.

[7] E. Commission. *Ethics guidelines for trustworthy AI*. Publications Office, 2019. DOI: doi/10.2759/346720.

[8] F. Ding et al. “Retiring Adult: New Datasets for Fair Machine Learning”. In: *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*. Ed. by M. Ranzato et al. 2021, pp. 6478–6490. URL: https://proceedings.neurips.cc/paper/2021/hash/32e54441e6382a7fbacbbbaf3c450059-Abstract.html.

[9] C. Düsing et al. “Towards predicting client benefit and contribution in federated learning from data imbalance”. In: *Proceedings of the 3rd International Workshop on Distributed Machine Learning*. 2022, pp. 23–29.

[10] M. Fontana et al. “Monitoring fairness in HOLDA”. In: *HHAI2022: Augmenting Human Intellect*. IOS Press, 2022, pp. 246–248.

[11] E. group on how AI principles should be implemented. “AI Governance in Japan”. In: (2023).

[12] Q. Li et al. “A Survey on Federated Learning Systems: Vision, Hype and Reality for Data Privacy and Protection”. In: *arXiv e-prints* (2019).

[13] T. Madiega. “Artificial intelligence act”. In: *European Parliament: European Parliamentary Research Service* (2021).

[14] B. McMahan et al. “Communication-Efficient Learning of Deep Networks from Decentralized Data”. In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*. Ed. by A. Singh et al. Vol. 54. Proceedings of Machine Learning Research. PMLR, 2017, pp. 1273–1282. URL: http://proceedings.mlr.press/v54/mcmahan17a.html.

[15] N. Ponomareva et al. “How to dp-fy ml: A practical guide to machine learning with differential privacy”. In: *Journal of Artificial Intelligence Research* 77 (2023), pp. 1113–1201.

[16] H. Roberts et al. “The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation”. In: *AI & society* (2021).

[17] T. Salazar et al. *A Survey on Group Fairness in Federated Learning: Challenges, Taxonomy of Solutions and Directions for Future Research*. 2024. URL: https://arxiv.org/abs/2410.03855.

[18] T. Yu et al. “Salvaging federated learning by local adaptation”. In: *ArXiv preprint* abs/2002.04758 (2020). URL: https://arxiv.org/abs/2002.04758.