Università di Genova
DIBRIS DIPARTIMENTO
DI INFORMATICA, BIOINGEGNERIA,
ROBOTICA E INGEGNERIA DEI SISTEMI

Fairness-Aware Federated Learning: A Comparative Study of Bias Mitigation Algorithms

Leonardo Gonfiantini

Master Thesis

Università di Genova, DIBRIS Via Opera Pia, 13 16145 Genova, Italy
https://www.dibris.unige.it/

Università
di Genova

MSc Computer Science
Data Science and Engineering Curriculum

Fairness-Aware Federated Learning: A Comparative Study of Bias Mitigation Algorithms

Leonardo Gonfiantini

Advisor: Barbara Catania

Examiner: Giovanna Guerrini

March, 2026

Abstract

Federated Learning (FL) is a distributed machine learning paradigm that enables multiple clients to collaboratively train a shared model while keeping data locally stored, thus preserving privacy. Despite its advantages, FL systems can inherit and amplify biases present in local datasets, potentially leading to unfair treatment of protected groups defined by sensitive attributes such as gender, race, or age. Such bias may arise from the learning process itself (*algorithmic bias*) or from imbalanced data distributions (*representation bias*), when certain groups are underrepresented.

This thesis investigates both forms of bias in FL through a comparative study of existing approaches and the introduction of a novel aggregation strategy. We consider three methods: FedAvg, the standard federated baseline; FairFed, a fairness-aware aggregation method targeting algorithmic bias; and FedCvg, which addresses representation bias by prioritizing clients whose local datasets satisfy predefined coverage constraints for protected groups. Since coverage thresholds are often domain-dependent and difficult to calibrate, we propose FedCvg-Ratio, a variant that adaptively assigns aggregation weights based on the deviation between local and global protected-group ratios.

All methods are implemented using *Flower*, an open-source federated learning framework, and evaluated on the Adult and COMPAS datasets. Results show that the fairness–accuracy trade-off is not universal: in several scenarios, FedCvg-Ratio combined with *local debiasing* improves both fairness and predictive performance. Overall, integrating client selection strategies such as *parity sampling* and *local debiasing* with fairness-aware aggregation yields the most consistent and robust fairness improvements in federated learning.

Acknowledgments

I would like to express my sincere gratitude to all the people who supported me throughout the development of this Master's thesis and during my entire Master's degree journey.

First of all, I would like to thank Barbara for the significant work we carried out together. Her support, guidance, and the time she dedicated to reviewing and correcting my thesis were invaluable throughout this process.

I am also deeply grateful to my family for their patience and support during this intense period. In particular, I would like to thank Stefano, Paola, and Andrea for always being there and for putting up with me during the demanding phases of this work.

A special thanks goes to my company, Gter, which gave me the opportunity to pursue my Master's degree while continuing to work. Their flexibility and trust made it possible for me to complete both my studies and this thesis.

Finally, I would like to thank my friends—too many to list individually—who constantly encourage me and with whom I share many passions and interests. Their support and friendship have always been a great source of motivation.

To all of you, thank you.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT-5.2 and Claude-Sonnet-4.5, Grammarly in order to: grammar and spelling check, paraphrase and reword. After using these tools/services, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

Table of Contents

List of Figures

List of Tables

Introduction

In recent years, the growing use of automated decision-making systems has raised significant concerns about fairness and bias [MMS+21]. Although these systems are often perceived as impartial and objective, they can produce unfair predictions for individuals or specific groups. This discrimination arises from societal biases that are reflected in, and possibly amplified by, data analytics pipelines [SARS23]. Understanding the impact of data-driven decisions at the social level and taking responsibility for them has become essential: we must ensure these systems benefit everyone without reinforcing discrimination or inequities.

From this perspective, data scientists and engineers play a crucial role: diversity, equity, and inclusion are increasingly recognized as fundamental quality dimensions throughout the entire data processing pipeline. The groups requiring protection are typically defined in terms of sensitive attributes—characteristics such as gender, race, religion, and economic status—depending on the specific context and applicable regulations. Often, these groups are defined using an intersectionality approach, combining multiple human-related properties to capture the complexity of real-world discrimination.

Bias in machine learning systems can manifest in two primary forms. *Representation bias* [SLAJ23] emerges during data preparation when certain subgroups are underrepresented in the training data, leading to insufficient information for accurate predictions on those groups. *Algorithmic bias* [MMS+21] concerns the quality of predictions made by automated decision-making systems and manifests when certain groups receive systematically unequal treatment, even when data representation is adequate.

A promising direction in machine learning is Federated Learning (FL) [MMR+17], a distributed machine learning paradigm in which clients collaborate to learn a global model over multiple rounds while keeping their data decentralized. At the beginning of each round, a subset of clients is sampled to perform model training on local datasets; only the updated local models are sent to a central server and aggregated to compute new model weights. These updated weights are then broadcasted back to the clients for subsequent training rounds. This method offers significant privacy and data security advantages due to the decentralized nature of the data, guaranteeing more responsible data usage [KMA+21].

Unlike traditional centralized machine learning, where data must be shared with the server, Federated Learning ensures that clients retain complete control over their data.

Building on these inherent advantages, increasing research effort is being devoted to harnessing Federated Learning for developing fair and accessible decision-making systems [ZCL22]. This includes strategies for handling data heterogeneity [LSZ+20] and ensuring equitable resource allocation across clients [LSBS19], as well as approaches that enhance fairness by incorporating algorithmic fairness constraints into Federated Learning algorithms. However, despite the growing adoption of fairness-aware solutions, limited techniques have been proposed to address representation bias in Federated Learning, even though its mitigation plays an essential role in ensuring the production of fair decision-making systems.

This thesis investigates both forms of bias in Federated Learning through a comparative study of existing approaches and the introduction of a novel aggregation strategy. We consider three baseline algorithms: FedAvg [MMR+17], the standard federated baseline; FairFed [EYH+23], a fairness-aware aggregation method targeting algorithmic bias through dynamic client weighting based on fairness metrics; and FedCvg [Bro23], which addresses representation bias by prioritizing clients whose local datasets satisfy predefined coverage constraints for protected groups.

A key limitation of FedCvg is its reliance on fixed coverage thresholds, which are typically domain-dependent and difficult to calibrate. To address this, we propose FedCvg-Ratio, a novel variant that adaptively assigns aggregation weights based on the deviation between each client's local protected-group ratio and the global ratio. This approach eliminates the need for practitioners to specify pre-defined thresholds that may not generalize across datasets or training stages, while maintaining the core principle of prioritizing clients that help balance global representation.

All methods are implemented using *Flower* [BTM+20], an open-source federated learning framework, and evaluated through extensive experiments on the Adult and COMPAS datasets, widely adopted in fairness research. Our experimental analysis considers multiple dimensions: (i) server-side aggregation strategies (FedAvg, FairFed, FedCvg, FedCvg-Ratio), (ii) local pre-processing techniques for bias mitigation (*local debiasing*), (iii) client selection strategies aimed at balancing group representation (*parity sampling*), (iv) both IID and heterogeneous non-IID data distributions controlled by Dirichlet partitioning, and (v) multiple fairness metrics, including Equalized Odds and Statistical Parity, alongside accuracy-based measures.

Our results demonstrate that the fairness–accuracy trade-off is not universal: in several scenarios, FedCvg-Ratio combined with local debiasing improves both fairness and predictive performance simultaneously. Algorithm effectiveness depends on data heterogeneity and group imbalance, with FairFed performing well in highly heterogeneous settings and coverage-based methods demonstrating robustness under severe representation disparities.

Moreover, local debiasing and parity sampling each contribute independently to fairness improvements, while their integration with fairness-aware aggregation yields the most consistent and robust results.

The remainder of this thesis is organized as follows:

- *Chapter 1* establishes the contextual foundation, introducing bias awareness, fairness metrics, and the Federated Learning paradigm.
- *Chapter 2* reviews existing approaches to fairness in Federated Learning, examining both algorithmic fairness and representation bias mitigation techniques, and positions our contributions within this landscape.
- *Chapter 3* presents the detailed algorithm descriptions, including the mathematical formulations and the intuition behind each approach. This chapter introduces our novel FedCvg-Ratio algorithm and discusses its advantages over existing methods.
- *Chapter 4* describes the experimental framework, including the Flower implementation, datasets, data partitioning strategies, and evaluation metrics.
- *Chapter 5* presents the experimental results, comparing algorithm performance across different heterogeneity levels and datasets.
- *Chapter 6* discusses the key findings, practical implications, limitations of our work, and concludes with directions for future research.

Chapter 1

Background

Data-driven automated decision systems are becoming increasingly prevalent in areas that directly affect people's lives—healthcare, autonomous driving, credit scoring, and criminal sentencing, to name a few. In recent years, a new distributed machine learning approach called Federated Learning has emerged, offering significant privacy and data security benefits that make it particularly suitable for training decision-making systems involving human-related data.

As these systems become more widespread, there has been a growing focus on evaluating the trustworthiness of their decisions. Significant research efforts now concentrate on reducing algorithmic bias and fostering fair decision-making models. However, bias is not an exclusive characteristic of automated decision systems; it can also develop from the data used to train these models. Bias can be an inherent dataset property, reflecting a systematic skew or favoritism within the data. This type of bias exists regardless of whether machine learning algorithms are employed and can result in an unfair or inaccurate representation of the underlying population. Specifically, a dataset may suffer from representation bias if it lacks sufficient information about specific subgroups.

This chapter establishes the contextual foundation for our study. Section 1.1 introduces the data analytics pipeline, providing a framework for understanding where different types of bias can emerge. Section 1.2 then defines bias, explores existing techniques to identify and reduce it in datasets, and discusses significant aspects of fairness investigation in machine learning. After surveying approaches proposed for centralized decision-making systems, we introduce the decentralized Federated Learning setting in Section 1.3.

1.1 The Data Analytics Pipeline

Before examining bias in detail, it is useful to understand the typical stages of a data analytics pipeline, as bias can emerge at different points throughout this process. Figure 1.1 illustrates a simplified view of the pipeline, from data collection to model deployment.

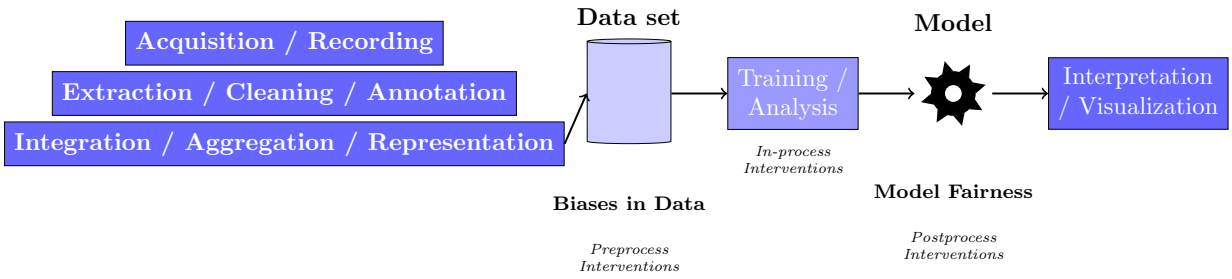


Figure 1.1: Data analytics pipeline showing stages from data collection to model deployment.

The data analytics pipeline consists of several stages through which data flows from collection to deployment. Understanding these stages is essential because different types of issues can emerge at different points in the process.

The pipeline begins with *data acquisition and recording*, where raw data is gathered from various sources such as sensors, surveys, or existing databases. This stage involves decisions about what data to collect, from whom, and how to record it.

Next comes *extraction, cleaning, and annotation*, where the raw data is processed to remove errors, handle missing values, and add labels or metadata. This preprocessing transforms raw data into a structured format suitable for analysis.

The *integration, aggregation, and representation* stage combines data from multiple sources, aggregates information at appropriate granularities, and chooses how to represent features (e.g., encoding categorical variables, normalizing numerical values).

Once the dataset is prepared, the *training and analysis* stage applies various analytical techniques to extract insights from the data. This includes machine learning algorithms for predictive modeling, data mining techniques for pattern discovery, OLAP (Online Analytical Processing) tasks for multidimensional analysis, and statistical analysis methods. These diverse analytical approaches enable different types of insights, from predictions on new data to exploratory analysis of complex relationships.

Finally, the *interpretation and visualization* stage involves understanding model outputs, evaluating performance, and presenting results to stakeholders. This stage is crucial for identifying potential issues before deployment.

Understanding this pipeline provides a framework for identifying where different types of problems can emerge and where interventions can be applied to address them.

1.2 Bias Awareness

Bias generally refers to a systematic tendency to favor or prejudice against certain groups or viewpoints, often without fair judgment. In the context of machine learning, Mehrabi et al. [MMS+21] survey how such biases manifest in data, algorithms, and user interactions. This definition captures the essence of the different types of bias people exhibit daily, whether consciously or unconsciously. It is nearly impossible for humans to be completely unbiased—forming judgments based on first impressions and preconceived notions is an instinctive behavior. Consequently, data produced and manipulated by humans can often be incomplete or inaccurate, reflecting these inherent biases.

As the definition highlights, bias typically manifests as a systematic disadvantage experienced by specific individuals or groups with particular characteristics. In some cases, discriminatory behavior targets specific individuals regardless of their group membership. More commonly, however, discrimination affects groups of people sharing similar traits that can be identified through a *sensitive* or *protected attribute* [BS16]. Examples of sensitive attributes include age, sex, and race—characteristics where underrepresented or disadvantaged groups are typically identifiable.

Referring back to the data analytics pipeline (Figure 1.1), we can identify two fundamental types of bias that emerge at different stages:

- *Representation bias* arises during the early stages (data collection and preprocessing) when certain groups are underrepresented in the dataset.
- *Algorithmic bias* manifests during the later stages (model training and evaluation) when the learning process produces unfair predictions for certain groups.

1.2.1 Representation Bias

Representation bias arises when a dataset underrepresents parts of the target population, independently of how the dataset is later used. This corresponds to the early stages of the data analytics pipeline (Figure 1.1): data acquisition, extraction, and integration.

As an example, consider a dataset used to train a medical diagnosis model for skin conditions. If the training images predominantly feature lighter skin tones, the model may perform poorly when diagnosing patients with darker skin tones simply because it has seen

few examples from this population during training. This underrepresentation exists in the dataset itself, before any learning algorithm is applied.

The reasons behind representation bias are various [SG19]: it can originate from incorrect data collection or be caused by biases introduced after collection, historically, cognitively, or statistically. Three primary causes can be identified:

1. *Historical bias*: Under-representation can arise from historical bias, defined as "the already existing bias and socio-technical issues in the world" [MMS+21]. Historical inequalities in society are reflected in the data collected about that society.
2. *Sampling and selection bias*: Selection bias occurs when there is no proper randomization when selecting people or groups to produce human-related data [OCDK19]. It is a cause of sampling bias, which happens when the gathered samples do not represent the population they aim to symbolize.
3. *Skewness of underlying distribution*: Representation bias may be caused by the skewness of the underlying distribution, which may lack sufficient representation for all subpopulations.

1.2.1.1 Identification of Representation Bias

Considering a sensitive attribute of a structured dataset, we can measure representation bias using metrics such as *representation rate* and *data coverage* [SARS23]:

- *Representation rate*: For a dataset D with a binary sensitive attribute $A \in \{0, 1\}$, the representation rate of group g is defined as:

$$RR_g = \frac{|D_g|}{|D|} \tag{1.1}$$

  where $|D_g|$ is the number of samples belonging to group g and $|D|$ is the total dataset size. A balanced dataset has $RR_0 \approx RR_1 \approx 0.5$. The more similar each group's representation rate, the less biased the data are with respect to that attribute.

  Beyond measurement, representation rate can be formulated as a constraint to ensure adequate representation. For a minimum threshold τ_{RR}, we require:

$$RR_g \geq \tau_{RR} \quad \forall g \tag{1.2}$$

  This constraint ensures that each group maintains at least a minimum proportion of the dataset, preventing severe underrepresentation that could lead to biased outcomes.

- *Data coverage*: Data coverage requires a minimum count for a specific group independent of the relative proportions. For a coverage threshold τ_C, we say that group g has adequate coverage if:

$$|D_g| \geq \tau_C \tag{1.3}$$

  This ensures that group g has sufficient samples for the learning algorithm to generalize, regardless of the overall distribution. Unlike representation rate which considers relative proportions, data coverage focuses on absolute counts to guarantee that each meaningful subgroup has enough instances for reliable analysis.

A high representation rate decreases issues related to representation bias in machine learning procedures; however, it is harder to achieve, especially if the underlying distribution is skewed. In that case, adequate coverage for each meaningful subgroup is more manageable and essential to ensure they are adequately represented.

1.2.1.2 Mitigation of Representation Bias

Representation bias mitigation focuses on addressing underrepresentation during the data collection and preprocessing stages, before any learning algorithm is applied. Based on recent surveys [SLAJ23, Acc23], mitigation strategies can be organized according to their location in the data pipeline and the specific techniques employed.

Data acquisition. The most effective way to address underrepresentation issues during data collection is to enrich the input dataset with additional samples through a data repair approach. However, data collection is typically costly. When data are acquired from third parties, monetary payments may be required; when collected directly, acquisition costs must still be considered. In both cases, further costs arise from cleaning, storing, and indexing the data. Therefore, it is crucial to identify the minimum number of samples that must be added to satisfy representation constraints. This problem was first investigated in [AJJ19], where efficient techniques were proposed to determine the least amount of additional data required to guarantee coverage with respect to multiple sensitive attributes. An efficient approach for coverage analysis over multiple relations is presented in [LGAJ20]. Early proposals were limited to categorical attributes with low cardinality, while the coverage-based data repair problem was later extended to ordinal and continuous-valued attributes in [ASJJ21]. When acquiring more real data is not feasible, synthetic data generation techniques can be employed to increase representation of minority groups, though care must be taken to avoid introducing artifacts.

Data transformation and preprocessing. Representation bias can also be introduced by data transformations defined in terms of selection-based queries, when the selected subset (i.e., the query result) does not ensure adequate group representation. In this setting, mitigation techniques correspond to query rewriting approaches that, given a query and

constraints expressed in terms of coverage or representation rate, search for the closest query to the original one that produces a result satisfying the desired constraints. Techniques for coverage-aware data transformations are presented in [AMC20, Acc23], while a related problem formulated in terms of representation rate is addressed in [SSAD22].

Data integration. Representation bias may also arise during data integration. An approach for determining, in a cost-effective manner, which additional data should be acquired for integration when distribution requirements expressed as coverage constraints are not met is discussed in [NAJ22].

1.2.2 Algorithmic Bias

Algorithmic bias manifests when the outcome of a learning process is unfair for some groups of the target population. Unlike representation bias, algorithmic bias does not refer to the dataset itself but to how the dataset is used in a specific analytical process. This corresponds to the later stages of the data analytics pipeline (Figure 1.1): training, analysis, and interpretation.

As an example, consider a resume screening system trained on historical hiring decisions from a technology company. Even if the training data contains adequate representation of all demographic groups, the model might learn to systematically favor candidates from certain educational backgrounds if historical hiring decisions reflected such preferences. The algorithm "learns" patterns present in the historical decisions, potentially perpetuating discrimination even when the input features (resumes) appear neutral. This bias emerges from how the learning algorithm processes the data, not from underrepresentation in the dataset itself.

1.2.2.1 Identification: Fairness Metrics

Algorithmic bias can be identified using specific fairness metrics that determine whether the outcomes are homogeneously distributed for different individuals or groups in the dataset [VR18]. Fairness metrics can be classified into *individual fairness* and *group fairness* metrics:

- *Individual fairness*: Similar individuals should be similarly treated in a learning algorithm outcome [DHP+12].
- *Group fairness*: Members of different groups should be equally treated. This interpretation relies on the definition of protected groups/minorities regarding sensitive attributes.

Group-based fairness metrics quantify how much the outcome of a decision-making process is influenced by the sensitive attribute values. Many different metrics have been proposed; most refer to classification tasks and are computed starting from confusion matrix values [VR18].

In the following, we discuss three well-known group-based metrics. We denote with A a binary sensitive attribute with values in $\{0, 1\}$, with $\hat{Y}$ the predicted binary label, and with Y the true value. In all three cases, values closer to 0 indicate better fairness, and by representing the unprivileged group with 0, positive values indicate that the unprivileged group outperforms the privileged one in the outcome.

Statistical Parity Difference (SPD). Statistical Parity rewards the classifier for classifying each group as positive at the same rate [DHP+12]. A binary predictor is considered fair if the probability of classifying an element as positive or negative is the same for both groups:

$$SPD = P(\hat{Y} = 1|A = 0) - P(\hat{Y} = 1|A = 1) \tag{1.4}$$

Statistical Parity ensures demographic parity but has limitations: enforcing fairness at the group level can be unjust to individuals, as it may require the algorithm to disregard otherwise qualified people to achieve outcome independence.

Equal Opportunity Difference (EOD). Equal Opportunity evaluates the performance of a binary predictor, which is considered fair if the true positive rate (TPR) is independent of the sensitive attribute A [HPS16]. It ensures that positive labels are assigned fairly by demanding that the model gives equal chances for individuals from various groups to be accurately identified as positive cases:

$$EOD = P(\hat{Y} = 1|A = 0, Y = 1) - P(\hat{Y} = 1|A = 1, Y = 1) \tag{1.5}$$

This metric focuses specifically on true positive rates, which is particularly important in applications where false negatives have serious consequences (e.g., medical diagnosis, loan approval).

Average Odds Difference (AOD). The Equalized Odds metric matches the True Positive Rate and the False Positive Rate for different groups [HPS16]. It demands TPRs and TNRs to be equal across groups. The Average Odds Difference is a relaxed version that measures the average difference between the FPR and TPR for the unprivileged and privileged groups:

$$AOD = \frac{(FPR_{A=0} - FPR_{A=1}) + (TPR_{A=0} - TPR_{A=1})}{2} \tag{1.6}$$

where $FPR_{A=a} = P(\hat{Y} = 1|A = a, Y = 0)$ and $TPR_{A=a} = P(\hat{Y} = 1|A = a, Y = 1)$.

Accuracy Difference (ACC_DIFF). Accuracy Difference measures the disparity in prediction accuracy between unprivileged and privileged groups. A model is considered fair under this metric if it achieves similar accuracy across both groups:

$$\text{ACC\_DIFF} = \text{Accuracy}_{A=0} - \text{Accuracy}_{A=1} \tag{1.7}$$

where $\text{Accuracy}_{A=a} = P(\hat{Y} = Y|A = a)$ represents the proportion of correct predictions for group a. This metric provides a straightforward measure of performance disparity and is particularly useful when overall predictive quality matters equally for all groups. Unlike metrics that focus on specific error types (e.g., false positives or false negatives), accuracy difference captures the overall fairness of the model's predictions across groups.

1.2.2.2 Mitigation of Algorithmic Bias

Several approaches have been proposed to mitigate algorithmic bias at different stages of a data analytics pipeline. Based on the stage at which they are applied (see Figure 1.1), mitigation techniques can be classified as follows.

Pre-processing techniques. They are designed to address imbalanced or biased distributions of sensitive attributes in datasets to reduce bias before model training. These techniques originate from the rich literature on fairness in centralized machine learning and have been adapted to distributed settings like Federated Learning because they can be applied locally without requiring access to the entire dataset [KC12, CWV+17]. Common pre-processing approaches include:

- *Removing sensitive attributes*: Excluding sensitive attributes and correlated features from the training dataset to prevent the model from learning discriminatory patterns. However, this approach may be insufficient if other features are correlated with the sensitive attribute.
- *Relabeling*: Changing specific samples' labels to remove discrimination from the input data, though this requires careful consideration to avoid introducing new biases.
- *Sample reweighting*: Assigning different weights to training samples based on their demographic group membership to balance the contribution of different (A, Y) combinations, where A is the sensitive attribute and Y is the label. Two prominent approaches are:
  - *Kamiran & Calders Reweighting* [KC12]: Assigns weights to achieve statistical independence between the sensitive attribute and the label. For each (A, Y)

cell, the weight is $w(A, Y) = \frac{P(A) \cdot P(Y)}{P(A,Y)}$, where probabilities are estimated from the dataset. We describe this method in detail in Chapter 3 as it forms the basis for our local debiasing technique.

  - *Inverse Propensity Weighting*: Assigns weights inversely proportional to the probability of observing each (A, Y) combination, upweighting rare combinations and downweighting common ones.

- *Data transformation*: Transforming the feature space to remove correlations between features and sensitive attributes. Learning Fair Representations [ZWS+13] learns an intermediate representation that preserves information about the target variable while removing information about the sensitive attribute. Fairness-aware feature selection identifies features less correlated with sensitive attributes, though this may sacrifice predictive accuracy.

When considering mitigation strategies during data preparation, several techniques have been proposed for different tasks.

Regarding data acquisition, the Slice Tuner framework [TW21] incorporates selective data acquisition strategies aimed at improving both fairness and accuracy across different data slices. Causal fairness has been considered for repairing datasets used to train classifiers in [SHS19]. In this case, the repaired dataset can be interpreted as a sample from a hypothetical fair world in which discriminatory causal relationships between sensitive attributes and outcomes have been removed.

If acquiring additional data is not feasible, data augmentation techniques can be employed to enhance the dataset. Data augmentation increases dataset size by generating synthetic samples or partially duplicating existing tuples. Oversampling is a common strategy, particularly useful when minority classes are underrepresented. Simple random oversampling may increase the risk of overfitting, whereas the Synthetic Minority Oversampling Technique (SMOTE) [CBHK02] generates synthetic minority instances based on their k-nearest neighbors, providing a more robust alternative.

Fairness-enhancing data cleaning interventions have been proposed in [TRO+19], where unfairness is mitigated during data sanitization using demographic parity as the reference nondiscrimination constraint. Mitigation approaches for data integration are discussed in [MPKF20], where associational and causal fairness are considered in automated data-wrangling pipelines to prevent downstream bias. Discriminatory bias in heterogeneous data integration is addressed in [YH21], which proposes EquiTensors, a learning-based approach that uses adversarial techniques to remove correlations with sensitive attributes. Finally, the impact of widely adopted data preparation procedures and the use of sensitive attributes on the fairness of machine learning systems is analyzed in [VLA19].

In-processing techniques. They aim to reduce bias during the learning procedure by

incorporating specific fairness constraints into training algorithms. A well-known approach is *Adversarial Debiasing* [ZLM18], which learns a classifier to maximize prediction accuracy while an adversary simultaneously tries to predict the sensitive attribute from the predictor's outputs. The predictor learns to make predictions that are accurate but uninformative about the sensitive attribute, effectively removing bias during training.

Post-processing techniques. They act on the model's output, leaving the learning phase untouched. These techniques usually involve changing the output labels to assign more positive labels to unprivileged groups, thus improving fairness indices [HPS16]. They require access only to the model predictions and information about the sensitive attributes, making them suitable for black-box contexts.

1.3 Federated Learning

Federated Learning (FL) is a machine learning paradigm introduced by McMahan et al. [MMR+17] in 2017 in their seminal paper "Communication-Efficient Learning of Deep Networks from Decentralized Data." The core idea is elegantly simple: learn a shared model by aggregating locally computed updates from distributed clients over a series of rounds, without ever centralizing the raw training data. Multiple clients collaborate to solve machine learning problems under the coordination of a central server, while their data remains decentralized on their own devices or systems.

The training process proceeds as follows: at the beginning of each round, the server broadcasts the current model weights to participating clients. Each client applies these weights, performs one round of training on its local dataset (which is never shared with the server), and transmits only the updated weights back to the server. The server then aggregates these updates and applies them to produce a new version of the global model, ready for the next round.

This architecture provides a distinct privacy advantage by decoupling model training from direct access to the raw training data. Moreover, the aggregation algorithm does not require knowledge of the source of each update, meaning that updates can be transmitted without identifying metadata. These properties make Federated Learning particularly attractive for applications involving sensitive personal data.

In the following, we introduce the core components and training process of Federated Learning (Section 1.3.1), describe the foundational FedAvg algorithm (Section 1.3.2), discuss the unique optimization challenges it presents (Section 1.3.3), and explore its applications across various domains (Section 1.3.4).

1.3.1 Introduction

The Federated Learning architecture consists of three main components:

1. *Server*: Coordinates the training process, maintains the global model, and aggregates client updates. The server never sees raw client data, only model parameters.
2. *Clients*: Local entities (e.g., mobile devices, hospitals, organizations) that possess private data and perform local training. Clients retain full control over their data.
3. *Communication protocol*: Defines how model parameters and metadata are exchanged between server and clients, typically optimized for bandwidth efficiency.

A typical FL training round proceeds through the following steps:

1. *Server broadcast*: The server sends the current global model θ^t to a subset of selected clients.
2. *Local training*: Each selected client k trains the model on its local dataset D_k for E local epochs, producing updated parameters θ_k.
3. *Client upload*: Clients send their updated model parameters back to the server.
4. *Server aggregation*: The server aggregates client updates to produce a new global model θ^{t+1}.
5. *Iteration*: Steps 1-4 repeat for T rounds until convergence or a stopping criterion is met.

Federated Learning algorithms aim to minimize the model's loss through the training rounds. In machine learning, loss is a key metric to evaluate models' performance by quantifying the difference between a given model's predicted and target labels. It is computed using a loss function, which may vary depending on the machine learning task and is used in both training and testing stages to improve and assess the quality of the model's predictions.

For binary classification tasks, a common loss function is *Binary Cross-Entropy*, evaluated as:

$$\text{BCE Loss} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i)] \tag{1.8}$$

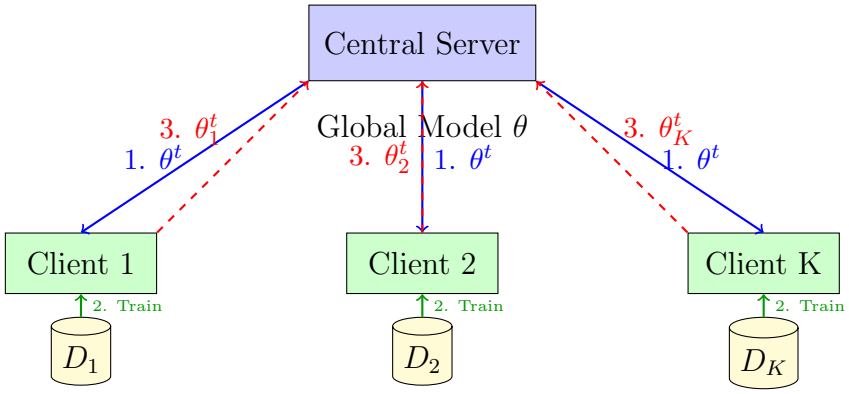


Figure 1.2: Federated Learning Architecture: The server broadcasts the global model to clients (step 1), clients train locally on their private data (step 2), and upload updated models (step 3). The server then aggregates updates to produce a new global model (step 4).

where N is the number of samples, y_i is the target label for the i-th sample, and $\hat{y}_i$ is the predicted probability.

Another valuable metric to evaluate a model's performance is **accuracy**, which measures how well a model performs on a classification task through the proportion of correct predictions out of the total predictions made:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \tag{1.9}$$

1.3.2 The FedAvg Algorithm

Federated Averaging, or FedAvg, is the foundational Federated Learning algorithm presented by McMahan et al. [MMR+17]. The algorithm operates through iterative rounds where the server broadcasts the current global model to a subset of selected clients, each client trains locally on its private data, and the server aggregates the updated models by computing a weighted average based on dataset sizes. This simple yet effective approach minimizes the global loss function—the weighted average of losses across all clients—while keeping data decentralized.

FedAvg serves as the baseline algorithm in our work and is described in detail in Section 3.2, where we present its mathematical formulation, algorithmic steps, and limitations that motivate fairness-aware extensions.

1.3.3 Federated Optimization Challenges

Federated Learning presents unique characteristics that differentiate it from classical distributed optimization problems. These aspects, collectively termed *Federated Optimization* [KMA+21], create both challenges and opportunities for algorithm design:

- *Non-IID data.* Perhaps the most significant challenge in Federated Learning is that data are non-Independent and Identically Distributed (non-IID). The training data on any given client typically do not represent the overall population's distribution, and the population distribution may not match that of any individual client. This heterogeneity is one of the primary obstacles to achieving good convergence and fairness in federated settings.

- Massive distribution. Federated Learning deployments are often massively distributed, with the number of clients potentially exceeding the average number of examples per client. This creates unique challenges for convergence analysis and communication efficiency that do not arise in traditional distributed computing.

- *Limited communication.* Communication between clients and the server may be constrained by bandwidth limitations, latency, or intermittent client availability. This necessitates algorithms that minimize the number of communication rounds while still achieving good model quality.

- *System heterogeneity.* Clients may possess vastly different computational capabilities, storage capacities, and network connectivity. This heterogeneity leads to varying training times and potential straggler effects, where slow clients delay the overall training process.

- *Privacy considerations.* While Federated Learning provides inherent privacy benefits by keeping data decentralized, additional considerations include ensuring the privacy of communications between clients and the server, protecting against inference attacks on model updates, and establishing trust among participants. A particular concern is *metadata disclosure*: even when raw data remains local, metadata such as dataset sizes, local fairness metrics, or demographic distributions can inadvertently reveal sensitive information about client populations. For instance, reporting the number of samples per sensitive group (as required by coverage-based algorithms) may expose the demographic composition of a hospital's patient base or an organization's employee diversity. This creates a tension between the need for aggregate statistics to ensure fairness and the risk of privacy leakage through metadata, requiring careful design of privacy-preserving aggregation mechanisms.

1.3.4 Applications

The capability of training models without accessing clients' raw data, combined with the massively distributed design, enables Federated Learning to be employed in contexts where other machine learning approaches suffer from data scarcity or privacy constraints [LSZ+20]. Moreover, various machine learning models—including linear models, random forests, and neural networks—can be trained in a Federated Learning setting.

These characteristics make Federated Learning an appropriate paradigm for building solutions across diverse domains:

- *Mobile applications*: The Google Keyboard input predictor (Gboard) was one of the first large-scale applications of FedAvg, enabling personalized next-word predictions without centralizing user typing data [HRM+18]. This demonstrated that Federated Learning could work at scale with millions of devices.
- *Healthcare*: Federated Learning enables collaborative disease prediction and diagnosis across hospitals without sharing sensitive patient data [RHL+20]. This addresses both stringent privacy regulations (such as HIPAA and GDPR) and the chronic data scarcity issues that plague medical AI research.
- *Finance*: Banks and financial institutions can collaboratively train fraud detection models without sharing customer transaction data, improving detection accuracy through collective learning while maintaining strict data privacy.
- *Internet of Things (IoT)*: Federated Learning has been introduced in wireless IoT systems to enhance communication efficiency and improve attack detection accuracy in smart home environments, where devices generate sensitive data that users prefer to keep local.

Human-related applications particularly benefit from Federated Learning's intrinsic privacy advantages. By allowing data to remain on users' devices, FL enables the collection of more personal data than users might otherwise consent to share, ultimately enabling better population representation while maintaining the data security guarantees that build user trust.

Chapter 2

Fairness in Federated Learning

The Federated Averaging algorithm presented in Section 1.3 does not explicitly consider fairness concerns, thus possibly producing a biased model. However, Federated Learning possesses an intrinsic fairness advantage stemming from its decentralized nature and resource allocation model. Unlike the centralized learning processes that dominate today's landscape—resource-intensive systems typically managed by corporations that often obscure essential information from users about data origins, transformations applied, and conclusions drawn—Federated Learning empowers individuals to manage their own data and the resources invested in the learning procedure. This represents a more sustainable, equitable, and responsible alternative to traditional approaches.

Building on these inherent advantages, extensive research has been conducted to identify and mitigate bias in Federated Learning processes, aiming to encourage fair decision-making models. This chapter reviews existing studies on this topic, examining different notions of fairness and complementary approaches to achieve it. We begin in Section 2.1 by clarifying the different meanings of fairness in the Federated Learning context. Section 2.2 then presents a taxonomy of approaches for group fairness in FL, which represents the focus of this work. Section 2.3 discusses the limited work on representation bias mitigation in FL. Finally, Section 2.4 synthesizes these approaches, discusses their limitations, and positions our contributions within this landscape.

2.1 Types of Fairness in Federated Learning

The notion of fairness in Federated Learning is multifaceted and can be interpreted from different perspectives, due to the distributed nature of the learning process [SACA26]. We can identify five main types of fairness in the FL context, described in the following.

Group Fairness (Demographic Fairness). This notion, inherited from centralized machine learning, focuses on ensuring that the global model treats different demographic groups equitably. Groups are defined based on sensitive attributes such as gender, race, or age. The goal is to ensure that model predictions do not discriminate against protected groups, as measured by fairness metrics like Equal Opportunity Difference (EOD), Statistical Parity Difference (SPD), or Average Odds Difference (AOD), introduced in Chapter 1.

Group fairness in FL faces unique challenges compared to centralized settings. The server lacks direct access to client data, making it difficult to apply standard pre-processing techniques. Moreover, data distributions across clients are typically non-IID and unbalanced, meaning that interventions improving global fairness metrics may not produce consistent results at the local level. This tension between global and local fairness represents one of the fundamental challenges in fair federated learning.

Individual Fairness. This notion of fairness requires that similar individuals receive similar outcomes from the machine learning model [DHP+12]. The principle is that the model should treat individuals with similar characteristics in a similar manner, regardless of their group membership. In the FL context, this could mean that two patients with similar medical histories across different hospitals (i.e., different clients) should receive similar predictions in a federated healthcare model, regardless of which hospital (client) their data belongs to.

Ensuring individual fairness in FL is particularly challenging because clients may have non-overlapping feature distributions, making it difficult to measure similarity across different datasets. Moreover, the decentralized nature of FL means that there is no global view of all individuals, complicating the enforcement of similarity-based fairness constraints. While individual fairness is an important notion, it is orthogonal to group fairness and is not the focus of this work.

Performance Distribution Fairness (Client Fairness). This notion, also known as client fairness, requires that the performance of the FL model (such as accuracy) is evenly distributed across all clients [LSBS19, MSS19, SACA26]. The goal is to prevent the global model from performing well on average while performing poorly for specific clients. This principle emphasizes the importance of uniformity in performance, ensuring that no single client is disproportionately advantaged or disadvantaged.

Performance distribution fairness can be formalized as minimizing the variance of model performance (e.g., accuracy or loss) across clients, or ensuring that no client experiences performance below a certain threshold. Note that performance distribution fairness is orthogonal to group fairness: a model can be fair at the client level (performing equally well for all clients) while still being unfair at the group level (discriminating against certain demographic groups).

Selection fairness. Also known as client participation fairness, this type focuses on the

fairness in selecting clients to participate in FL communication rounds [CWJ22, SACA26]. In each round of FL, a subset of clients is selected to update the global model. Selection fairness ensures that this process is unbiased and that all clients have an equitable opportunity to participate. This is important to prevent biases that could arise from consistently selecting certain clients over others.

For example, in a mobile keyboard FL system, consistently selecting high-end devices for training while excluding low-end ones can lead to biased predictions that favor users with expensive devices. In standard FL with random client selection, where a fixed fraction of clients is sampled uniformly at random each round, all clients have equal long-term selection probability. However, several factors can introduce selection bias:

- *Client availability*: Not all clients may be available in every round due to network connectivity, device status, or other practical constraints, leading to systematic underrepresentation of certain clients.
- *Strategic selection*: Techniques like Parity Sampling (discussed in Section 2.3) intentionally introduce selection bias to favor clients with better representation of minority groups. While this improves fairness, it means some clients may be selected more frequently than others.
- *System heterogeneity*: Clients with slower computation or communication may be excluded to avoid stragglers, potentially introducing bias if these clients have different data distributions.

Understanding and controlling client selection probabilities is crucial for both fairness and convergence. Strategic selection can improve representation balance but may reduce model diversity and potentially harm convergence on non-IID data.

Collaborative Fairness. Also known as contribution fairness, this notion focuses on rewarding clients proportionally to their contribution to the global model [LXWY20, SACA26]. The idea is that clients providing more valuable data or computational resources should receive better-performing personalized models or other incentives. This ensures that a client's reward is proportional to its contribution, which is important for motivating clients to actively participate and contribute with high-quality data.

For example, in a federated traffic prediction model, if a city contributing high-quality data receives the same reward as one providing noisy data, it discourages fair collaboration. Collaborative fairness raises interesting tensions with other fairness notions: rewarding high-contributing clients might disadvantage clients with smaller or lower-quality datasets, potentially exacerbating existing inequalities. Balancing collaborative fairness with group fairness and performance distribution fairness remains an open research challenge.

In this work, we focus primarily on *group fairness*, ensuring equitable treatment of demographic groups defined by sensitive attributes. This aligns with the broader societal goal of preventing discrimination in automated decision-making systems. However, we acknowledge that other fairness notions are important and may interact with group fairness in complex ways.

2.2 Group Fairness in Federated Learning: A Taxonomy

As discussed in Section 1.2.2, we can assess a machine learning algorithm's ability to avoid biases or discrimination by evaluating its algorithmic fairness. In the federated setting, however, this analysis becomes considerably more complex due to the intrinsic characteristics of the distributed environment.

The primary challenge stems from the lack of direct access to the datasets on which the model is trained, which prevents the application of standard pre-processing techniques commonly used in centralized settings. Furthermore, data distributions across clients are typically non-IID and unbalanced, meaning that interventions improving global fairness metrics may not produce consistent results at the local level when considering individual clients' datasets.

Several recent proposals address group fairness in Federated Learning, each taking a different approach to navigate these challenges. In the following, following the taxonomy proposed by Salazar et al. [SACA26], we classify these approaches based on *where* and *how* fairness interventions are applied and we highlight the main approaches further considered in this work.

Server-side aggregation modifications. These approaches modify the aggregation strategy at the server to incorporate fairness considerations. The key idea is to adjust the weights assigned to different clients' model updates based on fairness-related information. Two main types of approaches can be devised:

- *Fairness-based aggregation.* By adjusting the weight of each client's update based on fairness considerations, aggregation strategies can mitigate biases that might arise from uneven data distributions or varying levels of client participation FairFed [EYH+23] exemplifies this approach. At each round, clients evaluate the fairness of the current global model on their local data (before training) and report fairness metrics (e.g., EOD, SPD) to the server. The server computes a global fairness measure and calculates a "fairness gap" for each client—the deviation between the client's local fairness and the global target. Clients with smaller fairness gaps receive higher

aggregation weights, while those with larger gaps see their influence reduced. This approach is server-side and agnostic to the local debiasing method, allowing flexible use of different local debiasing techniques across clients. FairFed demonstrates significant improvements under high data heterogeneity but requires clients to share fairness-related information with the server.

- *Reweighted loss aggregation*: Rather than adjusting client aggregation weights, these approaches modify the weight of the local loss function for each sensitive group during training. FedFB [ZCL22] adapts the FairBatch algorithm [RLWS21] to FL. Clients calculate fairness metrics locally, which are safely aggregated at the server. After every k communication rounds, the server updates fairness coefficients—scalar weights λ_g for each sensitive group g that determine how much each group's loss contributes to the overall objective—to balance loss across different groups and broadcasts these coefficients back to clients. Clients use these coefficients to reweight samples during local training: samples from group g are weighted by λ_g in the loss function. While effective, this approach requires clients to share explicit information about model performance on each local subgroup, raising privacy concerns.

Client-side optimizations. These approaches modify the local training process at each client to incorporate fairness constraints, without requiring coordination with the server or other clients.

- *Pre-processing (local reweighting)*: Clients apply reweighting techniques (e.g., Kamiran & Calders [KC12], Inverse Propensity Weighting) during local training to balance the contribution of different demographic groups. This approach corresponds to standard pre-processing techniques from centralized ML (see Chapter 1) applied at the client side. Local reweighting is privacy-preserving (operates entirely on local data), modular (can be combined with any server-side aggregation strategy), and simple to implement. However, it has limitations: each client applies debiasing independently without coordination, which can lead to inconsistent effects; it addresses algorithmic bias but not representation bias; and it requires sufficient local diversity in each (A, Y) cell. In our work, we systematically evaluate the interaction between local reweighting and server-side aggregation strategies to understand when their combination is beneficial.

- *In-processing and post-processing techniques*: Clients can incorporate fairness constraints directly into their local optimization objectives (in-processing) or adjust model outputs after training (post-processing), as described in Chapter 1. These approaches train models that optimize both accuracy and fairness locally before sending updates to the server.

Hybrid approaches. These approaches combine server-side and client-side interventions to achieve better fairness outcomes. The majority of current fairness-aware FL solutions involve hybrid strategies where clients and the central server collaborate, leveraging the strengths of both local and global approaches while mitigating their respective weaknesses [SACA26].

Client participation approaches. These techniques involve strategically selecting or rewarding clients based on their contributions to fairness. Rather than modifying aggregation weights or local training, these approaches control which clients participate in each training round. By prioritizing clients whose data and updates enhance fairness across sensitive groups, the federated model can avoid disproportionately favoring or disadvantaging particular groups.

Zhang et al. [ZKW20] propose FairFL, a deep multi-agent reinforcement learning framework that optimizes both fairness and accuracy by training a client-selection policy function. Salazar et al. [SFAA23] introduce FAIR-FATE, which uses a validation set to evaluate and select specific clients whose local updates demonstrate higher fairness compared to the current global model. In our work, we implement and evaluate Parity Sampling [Bro23], a simpler approach that preferentially selects clients with better representation of underrepresented groups (discussed in detail in Section 2.3).

Constrained optimization approaches. Several works formulate fair FL as a constrained optimization problem, where fairness metrics are incorporated as constraints in the global objective [SACA26]. These approaches typically require each client to share statistics of sensitive attributes with the server, which can potentially leak information about local datasets. Examples include methods using Lagrangian multipliers, bi-level optimization, and multi-objective optimization frameworks to balance accuracy and fairness across clients.

Several patterns emerge from this taxonomy:

- *Privacy-fairness trade-off*: Most group fairness approaches require clients to share some fairness-related information with the server (fairness metrics, group statistics, or subgroup performance). This creates a tension between achieving fairness and maintaining privacy—one of the core motivations for using Federated Learning in the first place.
- *Heterogeneity challenge*: The effectiveness of fairness interventions depends heavily on the degree of data heterogeneity across clients. Methods that work well under mild heterogeneity may fail under extreme non-IID conditions, and vice versa.
- *Global vs. local fairness*: Interventions that improve global fairness metrics may not produce consistent fairness improvements at individual clients. This tension between global and local fairness remains a fundamental challenge.

- *Limited focus on representation bias*: The vast majority of existing work focuses on algorithmic fairness—ensuring that model predictions are fair given the available data. Very few works address representation bias—the underrepresentation of certain groups in the training data itself. This gap motivates our work, as we discuss in Section 2.4.

2.3 Representation Bias Mitigation in Federated Learning

While the approaches discussed in Section 2.2 focus on algorithmic fairness—ensuring equitable predictions across groups—a complementary and arguably more fundamental challenge in Federated Learning is *representation bias* (see Section 1.2.1).

This form of bias arises when certain groups are underrepresented in the combined training data across all clients, leading to models that perform poorly for minority groups regardless of the aggregation strategy employed.

Representation bias is particularly insidious in federated settings because the server has no direct visibility into the composition of client datasets. A client might possess abundant data overall but have very few examples from a protected minority group, and this imbalance propagates through the learning process. Addressing this challenge requires mechanisms that can infer and compensate for representation imbalances without violating the privacy guarantees that make Federated Learning attractive in the first place.

Despite its fundamental importance, representation bias in Federated Learning has received surprisingly little attention in the literature. To the best of our knowledge, only a handful of works explicitly address this challenge, and most focus on algorithmic fairness instead. This gap is significant because representation bias can undermine even the most sophisticated algorithmic fairness interventions—if certain groups are severely underrepresented in the training data, no amount of clever aggregation can fully compensate for the lack of information about those groups.

In the following, we discuss the limited work addressing representation bias in FL, focusing on coverage-based aggregation approaches and strategic client selection techniques.

Coverage-based approaches. The most relevant work addressing representation bias in FL is FedCvg, introduced in Brocchi's Master's thesis [Bro23]. FedCvg takes a fundamentally different approach from fairness-metric-based methods. Rather than adjusting weights based on model predictions (algorithmic fairness), FedCvg focuses on the *coverage* of protected groups in each client's dataset—that is, how well each client represents the minority population.

The central idea is to incorporate a coverage constraint (see Section 1.2.1 for the definition of data coverage)

directly into the aggregation process. Clients with better coverage of the protected group receive higher weights, ensuring that their contributions have greater influence on the global model. The coverage-based weight depends on the distance between the number of unprivileged samples a client possesses and a predefined coverage threshold.

This approach directly addresses representation bias by amplifying the voice of clients that better represent minority groups, effectively "repairing" the global data distribution without accessing individual data points. However, FedCvg requires practitioners to specify a coverage threshold, which may not generalize well across different datasets or training stages. The choice of this threshold can significantly impact the algorithm's effectiveness, and there is no principled way to determine the optimal value a priori.

Strategic client selection. Beyond modifying aggregation weights, another avenue for addressing representation bias lies in the client selection process itself. In traditional Federated Learning, clients are selected randomly at each round, but this approach can perpetuate representation imbalances: if most clients have skewed data distributions, random selection will consistently produce training batches that underrepresent minority groups.

Parity Sampling, also introduced in Brocchi's thesis [Bro23], exploits the insight that while we cannot select specific samples from a client's dataset (doing so would violate privacy guarantees), we *can* strategically choose which clients participate in each round. By favoring clients with better representation of underrepresented groups, we can effectively improve the global data distribution without accessing individual data points.

The Parity Sampling procedure operates as follows: in the first round, clients are selected randomly to establish baseline statistics about client compositions. From the second round onward, the server maintains a registry of each client's group composition and preferentially selects clients with more samples from the underrepresented group. A probability parameter p governs the selection strategy, allowing practitioners to tune the trade-off between fairness improvement and model diversity.

This approach is conceptually similar to client selection techniques like FAVOR [WKNL20] and Class Balancing [YWZ+20], but focuses on sensitive attribute distribution rather than class label distribution.

2.4 Gaps and Contributions

Having surveyed the landscape of fairness-aware federated learning approaches, we now synthesize the main gaps emerging in the literature and position our work within this

context.

The main gaps of the existing approaches for mitigating bias in federated learning can be summarized as follows:

Overwhelming focus on algorithmic fairness The vast majority of existing work focuses on algorithmic fairness, ensuring that model predictions are fair given the available data. Techniques like FairFed, FedFB, and local reweighting all operate under the assumption that sufficient data exists for all demographic groups, and the challenge is to ensure the learning algorithm treats these groups equitably.

However, this assumption often does not hold in practice. Federated Learning scenarios frequently involve highly heterogeneous data distributions where certain groups may be severely underrepresented across all clients. In such cases, algorithmic fairness interventions have limited effectiveness because the fundamental problem is lack of data, not biased learning.

Scarcity of representation bias solutions As highlighted in Section 2.3, very few works explicitly address representation bias in Federated Learning. To the best of our knowledge, only FedCvg and Parity Sampling [Bro23] directly tackle this challenge. This gap is particularly significant because:

- Representation bias is a root cause of unfairness that algorithmic interventions cannot fully address
- The distributed nature of FL makes representation bias more likely and harder to detect
- Addressing representation bias can amplify the effectiveness of algorithmic fairness techniques

Privacy-fairness trade-off Most group fairness approaches require clients to share fairness-related information with the server, creating tension with FL's privacy goals. For example, FairFed requires sharing fairness metrics (TP, FP, TN, FN counts per group), FedFB requires sharing subgroup performance statistics, and constrained optimization approaches require sharing sensitive attribute distributions. Coverage-based approaches like FedCvg and Parity Sampling require sharing group counts (n_{unpriv}, n_{total}).

All these metadata disclosures create potential privacy risks. While the shared information is aggregated and does not directly reveal individual data points, it can still leak sensitive information about client populations (e.g., demographic composition of a hospital's patients). Recent work on secure aggregation techniques, such as Per-element SecAgg [SKTH25], provides cryptographic mechanisms to protect against

data reconstruction attacks by ensuring that aggregated values are revealed only when sufficient clients contribute. However, applying such techniques to fairness-aware FL remains an open challenge, as fairness interventions often require fine-grained statistics that are difficult to compute under strong privacy constraints.

Fixed threshold limitations Coverage-based approaches like FedCvg require practitioners to specify a fixed coverage threshold before training begins. However, the optimal threshold depends on factors that may not be known in advance: the overall data distribution, the degree of heterogeneity across clients, and how these characteristics evolve during training. A threshold that works well for one dataset may perform poorly on another, limiting the generalizability of these approaches.

Limited understanding of complementary mechanisms While several works propose either server-side or client-side fairness interventions, few systematically investigate how these mechanisms interact when combined. Do server-side aggregation modifications and client-side local debiasing complement each other, or do they interfere? Under what conditions is their combination beneficial?

Building on these observations, our work makes the following contributions to address the identified gaps:

Comparative study of fairness-aware FL algorithms We implement and systematically evaluate three existing algorithms spanning different approaches to fairness in Federated Learning:

- *FedAvg* [MMR+17]: The standard FL algorithm without fairness considerations serves as our baseline, demonstrating the fairness issues that emerge when no mitigation is applied.
- *FairFed* [EYH+23]: We implement the fairness-aware aggregation algorithm, supporting multiple fairness metrics (EOD, SPD, Accuracy Difference) and investigating its behavior under various heterogeneity conditions.
- *FedCvg* [Bro23]: We implement the coverage-based aggregation algorithm that addresses representation bias by weighting clients based on their coverage of protected groups.

Note that Parity Sampling (strategic client selection) and local debiasing (sample reweighting) are orthogonal techniques that can be combined with any aggregation algorithm. In our experiments, we systematically evaluate all combinations to understand their interactions.

Design of a novel algorithm: FedCvg-Ratio As our main contribution, we introduce *FedCvg-Ratio*, a novel variant of FedCvg that addresses the key limitation of requiring a fixed coverage threshold. FedCvg-Ratio computes weights based on how each client's local ratio of protected samples differs from the global ratio, dynamically adapting to the current distribution. This approach eliminates the need for practitioners to specify a pre-defined threshold that may not generalize across datasets or training stages.

The key insight underlying FedCvg-Ratio is a shift in perspective: rather than asking "does this client have enough protected samples?" (a threshold-based question that requires specifying "enough"), we ask "does this client have more protected samples than average?" (a ratio-based question that adapts automatically). This relative comparison naturally adjusts as the global distribution evolves during training, making the algorithm more robust to varying data characteristics.

Additionally, FedCvg-Ratio incorporates Exponential Moving Average (EMA) smoothing to prevent abrupt weight changes between rounds that could destabilize training. The detailed algorithm description and mathematical formulation are presented in Chapter 3.

Systematic investigation of complementary mechanisms. We investigate the interaction between server-side fairness mechanisms (FairFed, FedCvg, FedCvg-Ratio) and client-side local debiasing techniques. By evaluating each algorithm both with and without local debiasing, we provide insights into:

- When do server-side and client-side interventions complement each other?
- Under what conditions does their combination provide additive benefits?
- Are there scenarios where combining both approaches is counterproductive?

This analysis provides practical guidance for practitioners seeking to deploy fairness-aware federated learning systems.

Focus on representation bias. Unlike most existing work that focuses on algorithmic fairness, our primary contribution addresses representation bias—the underrepresentation of certain groups in the training data. By developing FedCvg-Ratio and systematically evaluating coverage-based approaches, we advance the limited body of work on this fundamental challenge. Our experiments demonstrate that addressing representation bias can be more effective than algorithmic fairness interventions alone, particularly under extreme data heterogeneity.

Comprehensive experimental evaluation. We conduct extensive experiments across multiple datasets (Adult, COMPAS), varying levels of data heterogeneity (controlled by Dirichlet partitioning), and different sensitive attributes. This comprehensive

evaluation provides insights into how algorithm effectiveness depends on data characteristics, helping practitioners choose appropriate fairness interventions for their specific scenarios.

Table 2.1 summarizes the key characteristics of the bias-aware approaches we are going to consider in this work.

We selected FairFed as the representative of fairness-aware aggregation approaches (algorithmic bias mitigation) for several reasons: (1) it is well-documented with clear mathematical formulation and publicly available implementation; (2) it addresses algorithmic bias through adaptive client weighting based on fairness metrics, complementing our focus on representation bias; (3) it has been validated across multiple heterogeneity settings in the original paper; and (4) its server-side approach is orthogonal to client-side techniques, allowing us to study their interaction systematically.

According to the taxonomy presented in Section 2.2, in the table: *taxonomy category* points out the taxonomy node the considered algorithm belongs to; *level* refers to the component involved in the intervention (either client or server); *intervention type* refers to the approach used, consistent with the taxonomy categories; *metadata disclosure* refers to the specific metadata that clients communicate to the server beyond standard model weights; *bias type* indicates whether the approach addresses algorithmic bias or representation bias.

Table 2.1: Comparison of fairness-aware federated learning approaches

Algorithm	**Taxonomy category**	**Intervention type**	**Level**	**Metadata disclosed**	**Bias type**
FedAvg	Server-side aggregation	None (baseline)	Server	Dataset size (n)	-
FairFed	Server-side aggregation	Fairness-aware aggregation	Server	Fairness metrics (EOD, SPD, ACC_DIFF)	Algorithmic
FedCvg	Server-side aggregation	Coverage-based aggregation	Server	Group counts (n_{unpriv}, n)	Representation
FedCvg-Ratio	Server-side aggregation	Coverage-based aggregation	Server	Group counts (n_{unpriv}, n)	Representation
Local debiasing	Client-side optimization	Pre-processing (reweighting)	Client	None (local only)	Algorithmic
Parity Sampling	Client participation	Client participation	Server	Group counts (n_{unpriv}, n)	Representation

The following chapters present the detailed implementation of these algorithms (Chapter 3), the experimental framework (Chapter 4), and the systematic evaluation of their effectiveness (Chapter 5).

Chapter 3

Reference Algorithms

Building on the overview provided in Chapter 2, this chapter presents in detail the federated learning algorithms we implement and evaluate in this thesis. The chapter is organized as follows: Section 3.2 presents FedAvg, the baseline algorithm; Section 3.3 describes FairFed for fairness-aware aggregation; Section 3.4 covers FedCvg for representation bias mitigation based on coverage constraints; Section 3.5 introduces our novel contribution, FedCvg-Ratio, for representation bias mitigation based on representation-rate constraints; Section 3.6 describes Parity Sampling, a client selection strategy that can be combined with fairness-aware algorithms; and Section 3.7 describes the local debiasing technique that can be combined with any server-side algorithm. For each algorithm, we describe how it operates, present the mathematical formulation, and discuss limitations that motivate subsequent approaches.

3.1 Common Notation and Tasks

The algorithm presentation follows the notation presented in Table 3.1.

In all algorithms presented in this chapter, the *global model* θ^0 is initialized identically at the server. The server first loads the dataset to determine the input dimension, then creates either a linear model (logistic regression) or a multi-layer perceptron (MLP) based on configuration. The initial parameters are extracted from this freshly created model and used as θ^0. This ensures all algorithms start from the same baseline.

The individual algorithms also share the *client update procedure*. Algorithm 1 shows how each client performs local training. Each client i receives the current global model θ and performs local training using the ClientUpdate procedure (Algorithm 1, line 4). The client initializes its local model with the global parameters, then trains for E local epochs

Table 3.1: Notation for federated learning algorithms

Symbol	Description
R	Total number of training rounds
r	Current round index, $r \in \{1, \ldots, R\}$
C	Total number of available clients
S_r	Subset of clients selected for round r
i	Client index
D_i	Local dataset at client i
n_i	Number of training samples at client i, $n_i = \lvert D_i \rvert$
n_i^{unpriv}	Number of samples from the protected group (unprivileged, group 0) at client i
rr_i	Local ratio n_i^{unpriv}/n_i
rr_{global}^r	Global ratio across all participating clients in the current round: $\sum_i n_i^{unpriv} / \sum_i n_i$
θ^r	Global model parameters at round r
θ_i^r	Local model parameters at client i after round r
E	Number of local training epochs
η	Learning rate
B	Batch size
$\bar{w}_i$	Unnormalized (raw) aggregation weight for client i[1]
w_i	Normalized aggregation weight for client i[1]
ϕ_{global}	Aggregated global fairness metric
ϕ_i	Fairness metric for client i, computed on the global model before training
β	Fairness budget parameter that controls the speed of weight adjustment in FairFed
cov	Target coverage threshold (a fixed global parameter)
α	Sensitivity parameter that controls how coverage affect weights in FedCvg
λ	EMA smoothing factor
s_i	Ratio score for client i
p	Probability of using Parity Sampling w.r.t. random sampling

on its dataset D_i. During each epoch, the client processes its local dataset D_i in batches B, updating the model parameters using stochastic gradient descent with learning rate η:

$$\theta_i^r \leftarrow \theta_i^r - \eta \nabla \ell(\theta_i^r; B) \tag{3.1}$$

where $\ell(\theta_i^r; B)$ is the loss function evaluated on batch B, and $\nabla \ell(\theta_i^r; B)$ is its gradient with respect to the model parameters.

This procedure is referenced by all server-side algorithms (FedAvg, FairFed, FedCvg, FedCvg-Ratio).

[1]Superscript ^{r} is with this notation used to denote round-dependent weight values.

Additionally, at each round of the procedure a subset S_r of the set of clients C can be selected for the processing. The percentage of selected clients is usually called *fraction fit*. The selection could be random or based on other selection approaches. In all the server-side algorithm we use function ClientSelection(ff, C) to denote the function that selects $ff \cdot |C|$ clients from the whole set C and we assume this selection is randomly performed if not otherwise stated.

Algorithm 1 ClientUpdate(i, θ) (common to all algorithms)

$\theta_i \leftarrow \theta$
for epoch $e = 1$ to E **do**
    for batch $B \in D_i$ **do**
        $\theta_i \leftarrow \theta_i - \eta \nabla \ell(\theta_i; B)$
    end for
end for
return θ_i

3.2 FedAvg: Federated Averaging

FedAvg, introduced by McMahan et al. [MMR+17], serves as the foundational algorithm for Federated Learning and our baseline for comparison.

3.2.1 Introduction

FedAvg operates through a simple yet effective iterative process. At the beginning of each training round, the central server broadcasts the current global model to a selected subset of clients. Each client then trains this model on its local dataset for a specified number of epochs, producing an updated local model. The clients send their updated models back to the server, which combines them into a new global model by computing a weighted average, where each client's contribution is proportional to the size of its local dataset.

The intuition behind this weighting scheme is straightforward: clients with more data have seen more examples and thus their updates should carry more influence. This approach works well when the goal is purely to minimize the global loss function, but it does not account for fairness considerations—a client with a large but biased dataset will have substantial influence on the global model, potentially propagating its biases.

3.2.2 Description

The FedAvg algorithm proceeds as shown in Algorithm 2. We now explain each component in detail, using the notation introduced in Table 3.1.

The server initializes the global model θ^0 (line 2). Then, the algorithm operates in the following main steps at each round (R in total, lines 3-14):

1. *Client sampling* (line 4). The server selects a subset of clients S_r to be used in round r, based on the fraction fit ff.
2. *Local training and metadata collection* (lines 5-9). Each client $i \in C$ receives the current global model θ^{r-1}, performs local training using the ClientUpdate procedure (Algorithm 1, line 6), and get back the new model parameters for each of them (line 6). The server also collects some additional metadata from the client (line 7), in this case corresponding to the total number of client samples, used to inizialize the raw weights (line 8).
3. *Weight normalization* (lines 10-13). The client raw weights are normalized to sum to 1 (lines 10-13).
4. *Aggregation* (line 14). Finally, the new global model is computed as a weighted average of the client models (line 12). Since the weight w_i assigned to client i is proportional to its dataset size, this ensures that clients with more data have greater influence on the global model.

Using this algorithm, FedAvg minimizes the following global objective function:

$$\min_{\theta} f(\theta) = \sum_{i=1}^{C} \frac{n_i}{n} \cdot F_i(\theta) \tag{3.2}$$

where $f(\theta)$ is the global loss function, $F_i(\theta) = \frac{1}{n_i} \sum_{(x,y)\in D_i} \ell(\theta; x, y)$ is the local loss function at client i (with (x, y) representing a training sample with features x and label y contained in the client dataset D_i), and $n = \sum_{i=1}^{C} n_i$ is the total number of examples across all clients.

3.2.3 Limitations

FedAvg has several limitations regarding fairness that motivate the development of fairness-aware alternatives:

Algorithm 2 FedAvg

Server:
Initialize global model θ^0
for round $r = 1$ to R **do**
    $S_r = \text{ClientSampling}(ff, C)$
    for each client $i \in S_r$ **in parallel do**
        $\theta_i^r \leftarrow \text{ClientUpdate}(i, \theta^{r-1})$
        Collect n_i from client i
        $\bar{w}_i \leftarrow n_i$
    end for
    Compute $w^r = \sum_{i \in S_r} \bar{w}_i$
    for each client $i \in S_r$ **do**
        $w_i^r \leftarrow \bar{w}_i / w^r$
    end for
    $\theta^r \leftarrow \sum_{i \in S_r} w_i^r \cdot \theta_i^r$
end for

- *No fairness optimization*: FedAvg does not consider fairness metrics in its aggregation scheme. The algorithm optimizes solely for minimizing the global loss, without any mechanism to ensure equitable treatment of protected groups.
- *Bias amplification*: Clients with larger datasets have more influence on the global model, even if their data is more biased. This can amplify existing biases in the training data.
- *Representation bias propagation*: Under-represented groups may have little influence on the global model, as clients with few samples from minority groups contribute less to the aggregation.
- *No strategic selection*: Random client selection may consistently exclude clients with important minority representation, perpetuating representation imbalances.

These limitations establish the need for fairness-aware approaches that can maintain the benefits of federated learning while ensuring equitable treatment of all demographic groups.

3.3 FairFed: Fairness-Aware Aggregation

FairFed, proposed by Ezzeldin et al. [EYH+23], addresses the fairness limitations of FedAvg by dynamically adjusting client aggregation weights based on their fairness metrics.

3.3.1 Introduction

FairFed builds on a key insight: not all client updates should be treated equally when fairness is a concern. The algorithm works by first having each client evaluate the fairness of the current global model on its local data *before* performing any training. This evaluation produces a local fairness metric (such as SPD, EOD, or ACC_DIFF) that indicates how fairly the model treats different demographic groups from that client's perspective.

The server then aggregates these local fairness metrics to compute a global fairness measure. For each client, the algorithm calculates a "fairness gap"—the difference between the client's local fairness and the global fairness. Clients whose local fairness is closer to the global target receive higher aggregation weights, while clients with larger fairness gaps see their influence reduced.

The intuition is that clients whose data produces fairness metrics similar to the global target are more "aligned" with the overall fairness goal, and their model updates should carry more weight. Conversely, clients whose local fairness deviates significantly from the global target may have unusual data distributions that could harm global fairness if given too much influence.

3.3.2 Description

The FairFed algorithm proceeds as shown in Algorithm 3. We now explain each step in detail, using the notation introduced in Table 3.1.

The server initializes the global model θ^0 (line 2) and client weights $\bar{w}_i^0 = n_i / \sum_{j=1}^{C} n_j$ for all clients (line 3). This ensures that the algorithm starts from a reasonable baseline similar to FedAvg, as specified in the paper. Then, the algorithm operates in several steps at each round (R in total, lines 4-23):

1. *Client sampling* (line 5). The server selects a subset of clients S_r to be used in round r, based on the fraction fit ff.

2. *Pre-training fairness evaluation* (lines 6-8). Before local training begins, each client i evaluates the current global model θ^{r-1} on its local validation data to compute its local fairness metric ϕ_i (line 7). The algorithm supports any fairness metric introduced in Section 1.2.2 of Chapter 1. The specific metric used is a configuration parameter.

3. *Global fairness computation* (line 9). The server aggregates fairness metrics from all clients. Notice that in the algorithm, line 7 shows $\phi_i \leftarrow \text{EvaluateFairness}(i, \theta^{r-1})$ as an abstraction. In practice, this function returns both the local fairness metric

Algorithm 3 FairFed(β, ϕ)

Server:
Initialize global model θ^0
Initialize client weights $\bar{w}_i^0 = n_i / \sum_j^C n_j$ for all i
for round $r = 1$ to R **do**
    $S_r \leftarrow$ ClientSampling(ff, C)
    for each client $i \in S_r$ **in parallel do**
        $\phi_i \leftarrow$ EvaluateFairness(i, θ^{r-1}, ϕ)
    end for
    $\phi_{global} \leftarrow$ AggregateFairness($\{\phi_i | \forall i \in S_r\}$)
    $\bar{\Delta} \leftarrow$ mean($|\phi_{global} - \phi_i|$ for $i \in S_r$)
    for each client $i \in S_r$ **do**
        $\Delta_i \leftarrow |\phi_{global} - \phi_i|$
        $\bar{w}_i^r \leftarrow \max\{0, \bar{w}_i^{r-1} - \beta \cdot (\Delta_i - \bar{\Delta})\}$
    end for
    $w \leftarrow \sum_{i \in S_r} \bar{w}_i^r$
    for each client $i \in S_r$ **do**
        $w_i^r \leftarrow \bar{w}_i^r / w$
    end for
    for each client $i \in S_r$ **in parallel do**
        $\theta_i^r \leftarrow$ ClientUpdate(i, θ^{r-1})
    end for
    $\theta^r \leftarrow \sum_{i \in S_r} w_i^r \cdot \theta_i^r$
end for

ϕ_i and the raw counts needed for global aggregation (e.g., $TP_{A=0,i}$, $P_{A=0,i}$, $TP_{A=1,i}$, $P_{A=1,i}$ for EOD).

The server then aggregates these raw counts to compute the global fairness metric ϕ. For example, for EOD, the global True Positive Rates are computed as:

$$TPR_{A=0}^{global} = \frac{\sum_{i \in C} TP_{A=0,i}}{\sum_{i \in C} P_{A=0,i}}, \quad TPR_{A=1}^{global} = \frac{\sum_{i \in C} TP_{A=1,i}}{\sum_{i \in C} P_{A=1,i}} \tag{3.3}$$

This aggregation approach ensures that the global fairness metric accurately reflects the combined data distribution across all clients. Crucially, we aggregate the raw counts (TP, P) rather than averaging the local ϕ_i values. Averaging local metrics would give incorrect results when clients have different dataset sizes, as it would weight all clients equally regardless of their data volume. By aggregating raw counts, we ensure that the global metric is computed correctly over the entire federated dataset.

4. *Fairness gap calculation and weight update* (lines 10-14). First, the average fairness gap across all participating clients is computed as $\bar{\Delta} = \frac{1}{|S_r|} \sum_{i \in S_r} |\phi_{global} - \phi_i|$ (line 10). Then, for each client i, the fairness gap Δ_i measures how much its local fairness deviates from the global fairness, and the weight is updated accordingly (lines 11-14, including the end for at line 14).

   When the local fairness metric is undefined (e.g., when a client lacks positive examples from one protected group), FairFed uses an accuracy-based gap as fallback:

$$\Delta_i = |Acc_{global} - Acc_i| \tag{3.4}$$

   The weight update follows the paper's formula (Equation 6 in [EYH+23]), combining the gap calculation with the weight adjustment in a single loop (lines 11-14):

   The intuition behind this update is:

   - If $\Delta_i > \bar{\Delta}$ (client has worse fairness than average): weight decreases.
   - If $\Delta_i < \bar{\Delta}$ (client has better fairness than average): weight increases.
   - If $\Delta_i = \bar{\Delta}$ (client has average fairness): weight unchanged.

   The parameter β controls the speed of weight adjustment.

5. *Weight normalization* (lines 15-18). The unnormalized weights $\bar{w}_i^r$ are normalized to sum to 1 (line 15 computes the sum, lines 16-18 normalize):

6. *Local training and aggregation* (lines 19-22). After local training using ClientUpdate (Algorithm 1) (lines 19-21), the server aggregates client models using the fairness-aware weights (line 22).

3.3.3 Limitations

While FairFed represents a significant advance over FedAvg for fairness, it has several limitations:

- *Metric dependency*: The algorithm's effectiveness depends on the ability to compute meaningful fairness metrics at each client. Highly skewed local distributions can make metrics undefined.
- *Homogeneous settings*: As noted by the original authors, FairFed's advantages diminish in highly homogeneous settings where local debiasing alone may suffice.
- *No representation bias mitigation*: FairFed adjusts weights based on model fairness metrics, but does not directly address representation bias in the data distribution.

3.4 FedCvg: Coverage-Based Aggregation

FedCvg, introduced in Brocchi's Master's thesis [Bro23], addresses representation bias through coverage-based aggregation weights that favor clients with better coverage of protected groups.

3.4.1 Introduction

FedCvg takes a fundamentally different approach from FairFed. Instead of adjusting weights based on how the model performs (fairness metrics), FedCvg adjusts weights based on what data each client has (coverage of protected groups). The key insight is that representation bias—the underrepresentation of certain groups in the training data—is a root cause of unfair models, and addressing it directly at the data level can be more effective than trying to correct for it at the model level.

At each round, clients report how many samples they have from the protected (unprivileged) group. The server uses this information to compute coverage-based weights: clients with more samples from the protected group receive higher weights, ensuring their model updates have greater influence on the global model. This effectively "repairs" the global data distribution by giving more voice to clients that better represent minority groups.

Note that in FedCvg, the same clients tend to receive consistently higher weights across rounds because their dataset composition (number of protected group samples) remains constant. This differs from FairFed, where weights adapt based on model performance metrics that can change over time.

3.4.2 Description

The FedCvg algorithm proceeds as shown in Algorithm 4. We now explain each step in detail, using the notation introduced in Table 3.1.

The server initializes the global model θ^0 (line 2). Then, the algorithm operates in several steps at each round (R in total, lines 3-11):

1. *Client sampling* (line 4). The server selects a subset of clients S_r to be used in round r, based on the fraction fit ff.
2. *Local training and metadata collection* (lines 5-8). Each client $i \in S_r$ performs local training in parallel using ClientUpdate (Algorithm 1, line 6). Each client reports not only its updated model θ_i^r but also its dataset statistics: total samples n_i and number

Algorithm 4 FedCvg(α_{cov})

1: **Server:**
2: Initialize global model θ^0
3: **for** round $r = 1$ to R **do**
4: $S_r = \text{ClientSampling}(ff, C)$
5: **for** each client $i \in S_r$ **in parallel do**
6: $\theta_i^r \leftarrow \text{ClientUpdate}(i, \theta^{r-1})$
7: Collect n_i, n_i^{unpriv} from client i
8: **end for**
9: **for** each client $i \in S_r$ **do**
10: $\bar{w}_i \leftarrow \exp(\alpha_{cov} \cdot (n_i^{unpriv} - cov)) \cdot n_i$
11: **end for**
12: Compute $w^r = \sum_{i \in S_r} \bar{w}_i$
13: **for** each client $i \in S_r$ **do**
14: $w_i^r \leftarrow \bar{w}_i / w^r$
15: **end for**
16: $\theta^r \leftarrow \sum_{i \in C} w_i^r \cdot \theta_i^r$
17: **end for**

of samples from the protected group n_i^{unpriv} (line 7). This information is essential for computing coverage-based weights.

3. *Weight computation and normalization* (lines 9-16). For each client $i \in S_r$, the server computes the coverage weight (line 10), according to [Bro23]:

$$\bar{w}_i = \exp(\alpha_{cov} \cdot (n_i^{unpriv} - cov)) \cdot n_i \tag{3.5}$$

The weight is computed assuming that a domain dependent coverage constraint for the unprivileged group is available with a threshold *cov*. The formula relies on a parameter α_{cov} that weights the impact of the coverage on the weight computation. The intuition behind this formula is:

- If $n_i^{unpriv} > cov$: client i has good coverage of protected group → weight increases ($\exp(\alpha_{cov} \cdot \text{positive}) > 1$).
- If $n_i^{unpriv} < cov$: client i has poor coverage of protected group → weight decreases ($\exp(\alpha_{cov} \cdot \text{negative}) < 1$).
- If $n_i^{unpriv} = cov$: client i meets target → weight $= n$ (like FedAvg).

This weighting scheme ensures that clients with better representation of the protected (unprivileged) group have greater influence on the global model, effectively addressing

representation bias by amplifying the voice of clients that better represent minority groups.

The raw weights $\bar{w}_i$ are then normalized to sum to 1 (lines 13-16).

4. *Aggregation* (line 18). The server finally aggregates client models using the normalized coverage-based weights.

3.4.3 Limitations

FedCvg has several limitations that motivate our proposed FedCvg-Ratio:

- *Fixed threshold*: The coverage threshold *cov* must be specified a priori and may not generalize across datasets or training stages. There is no principled way to determine the optimal value.
- *Absolute vs. relative*: The threshold is an absolute number of samples, which does not adapt to varying dataset sizes. A $cov = 100$ may be appropriate for large clients but too demanding for small clients.
- *Static weighting*: Since client dataset compositions don't change, the same clients receive consistently high weights throughout training, potentially limiting model diversity.
- *Parameter sensitivity*: The effectiveness depends heavily on the choice of both *cov* and α, requiring careful tuning for each dataset.

These limitations motivate our novel contribution, FedCvg-Ratio, which addresses them through a ratio-based approach, following the notion of representation rate introduced in Section 1.2.1.

3.5 FedCvg-Ratio: a New Algorithm based on Representation Rate

FedCvg-Ratio is our novel contribution to fairness-aware federated learning. It addresses the key limitations of FedCvg by replacing the fixed coverage threshold with a dynamic, ratio-based weighting scheme that adapts to the current data distribution.

3.5.1 Introduction

The primary limitation of FedCvg is its reliance on a fixed coverage threshold *cov*. This threshold must be specified before training begins, but the optimal value depends on factors that may not be known in advance: the overall data distribution, the degree of heterogeneity across clients, and how these characteristics evolve during training. A threshold that works well for one dataset may perform poorly on another, and a threshold that is appropriate at the beginning of training may become suboptimal as the model converges.

FedCvg-Ratio addresses this limitation through a fundamental shift in perspective. Rather than asking "does this client have enough protected samples?" (a threshold-based question that requires specifying "enough" via the fixed *cov* parameter), we ask "does this client have more protected samples than average?" (a ratio-based question that adapts automatically to the current round's distribution). This relative comparison naturally adjusts as the global distribution evolves during training, eliminating the need for practitioners to specify a fixed threshold.

FedCvg-Ratio thus directly operationalizes the Representation Rate concept from Section 1.2.1. Recall that for a client i, the local representation rate of the unprivileged group is:

$$rr_i = \frac{n_i^{unpriv}}{n_i} \tag{3.6}$$

FedCvg-Ratio compares each client's local ratio rr_i, the local ratio) against the global representation rate $rr_{global} = \frac{\sum_i n_i^{unpriv}}{\sum_i n_i}$, and adjusts weights accordingly. This makes the algorithm inherently adaptive: as r_{global} changes across rounds (due to different client participation), the weighting automatically adjusts.

The key insight is that this comparison is inherently adaptive. If the global ratio changes during training (for example, as different clients participate in different rounds due to Parity Sampling), the algorithm automatically adjusts its weighting scheme using a *ratio score*. A client that was considered "above average" in one round might be "below average" in another, depending on which other clients participated. By dinamically adapting the ratio score makes the algorithm more robust to varying data characteristics.

Example 3.1 *Consider a scenario with $rr_{global} = 0.33$ (33% unprivileged, 67% privileged). Table 3.2 shows data distribution information for three clients A, B, C. Client B, with 50% unprivileged samples, is above the global average of 33%, so it receives a boost (ratio score s_i greater than 1). Client A, with only 20%, is below average and receives a penalty (ratio score s_i lower than 1). Client C matches the global average and receives neutral weighting*

(ratio score s_i equal to 1). The result is that the aggregated model "sees" a more balanced distribution because clients with more unprivileged samples have higher weight.

To further enhance stability, FedCvg-Ratio incorporates Exponential Moving Average (EMA) smoothing [Hun86]. Rather than using the raw computed weights directly, the algorithm maintains a smoothed version that blends the current weights with historical values. This prevents abrupt weight changes between rounds that could destabilize training, particularly in early rounds when the global distribution estimate may be noisy.

Table 3.2: FedCvg-Ratio weight computation example

Client	n_i	n_i^{unpriv}	rr_i	s_i	**Effect**
A	1000	200	0.20	∼0.80	Penalized (below average)
B	1000	500	0.50	∼1.25	Boosted (above average)
C	1000	330	0.33	∼1.00	Neutral (matches global)

3.5.2 Description

The FedCvg-Ratio algorithm proceeds as shown in Algorithm 5. We now explain each step in detail, using the notation introduced in Table 3.1.

At the beginning, the server initializes the global model θ^0 (line 2) and EMA weights $w_i^0 = n_i / \sum_j n_j$ for all clients (line 3). The algorithm then operates in several steps at each round (R in total, lines 4-41):

1. *Client sampling* (line 5). The server selects a subset of clients S_r to be used in round r, based on the fraction fit ff.

2. *Local training and metadata collection* (lines 6-9). Each client $i \in S_r$ performs local training in parallel using ClientUpdate (Algorithm 1, line 6). The server collects both the updated model θ_i^r and dataset statistics: total samples n_i and number of samples from the protected group n_i^{unpriv} (line 8).

3. *Global statistics* (lines 10-12). The server computes the total samples and unprivileged samples across all participating clients in the current round (lines 10-11), then calculates the global ratio rr_{global}^r (line 12). Note that rr_{global}^r is computed only over the clients participating in the current round r, not over all clients in the federation. This makes rr_{global}^r dynamic: it can change from round to round as different clients participate, especially when combined with Parity Sampling.

Algorithm 5 FedCvg-Ratio(α_{rr}, λ)

Server:
Initialize global model θ^0
Initialize EMA weights $w_i^0 = n_i / \sum_j n_j$ for all clients
for round $r = 1$ to R **do**
    $S_r = \text{ClientSampling}(ff, C)$
    for each client $i \in S_r$ **in parallel do**
        $\theta_i^r \leftarrow \text{ClientUpdate}(i, \theta^{r-1})$
        Collect n_i, n_i^{unpriv} from client i
    end for
    $n \leftarrow \sum_{i \in S_r} n_i$
    $n^{unpriv} \leftarrow \sum_{i \in S_r} n_i^{unpriv}$
    $rr_{global}^r \leftarrow n^{unpriv} / n$

    for each client $i \in S_r$ **do**
        $rr_i \leftarrow n_i^{unpriv} / n_i$
        $diff \leftarrow rr_i - rr_{global}^r$
        $norm\_diff \leftarrow diff / \min(rr_{global}^r, 1 - rr_{global}^r)$

        if $rr_{global}^r < 0.5$ **then**
            $s_i \leftarrow 1 + \alpha_{rr} \cdot norm\_diff$
        else
            $s_i \leftarrow 1 - \alpha_{rr} \cdot norm\_diff$
        end if
        $s_i \leftarrow \max(0.5, \min(2.0, s_i))$

        $\bar{w}_i^r \leftarrow n_i \cdot s_i$
    end for

    Compute $w^r = \sum_{i \in S_r} \bar{w}_i^r$
    for each client $i \in S_r$ **do**
        $\hat{w}_i^r \leftarrow \bar{w}_i^r / w^r$
    end for

    for each client $i \in S_r$ **do**
        if client i seen before **then**
            $w_i^r \leftarrow \lambda \cdot w_i^{r-1} + (1 - \lambda) \cdot \hat{w}_i^r$
        else
            $w_i^r \leftarrow \hat{w}_i^r$
        end if
    end for

    $\theta^r \leftarrow \sum_{i \in S_r} w_i^r \cdot \theta_i^r$
end for

4. *Ratio computation* (lines 14-17). For each client $i \in S_5$, we compute the following ratio-based scores:

   - Local ratio: $rr_i = n_i^{unpriv}/n_i$ (line 15)
   - Difference from global: $diff = rr_i - rr^r_{global}$ (line 16)
   - Normalized difference: $norm\_diff = diff / \min(rr^r_{global}, 1 - rr^r_{global})$ (line 17)

   The normalization ensures consistent effect regardless of how imbalanced the global distribution is. A client that is 10% above average receives the same relative boost whether $rr^r_{global} = 0.3$ or $rr^r_{global} = 0.7$.

5. *Ratio score and weight update* (lines 19-26). The ratio score s_i adjusts the client's weight based on whether it helps balance the distribution (lines 19-23), then is clamped to prevent extreme values (line 24), and finally the raw weight is computed (line 26). The definition relies on a parameter α_{rr} to tune the impact of $norm\_diff$ on the score as follows:

$$s_i = \begin{cases} 1 + \alpha_{rr} \cdot norm\_diff & \text{if } rr^r_{global} < 0.5 \text{ (need more unprivileged)} \\ 1 - \alpha_{rr} \cdot norm\_diff & \text{if } rr^r_{global} \geq 0.5 \text{ (need more privileged)} \end{cases} \tag{3.7}$$

   The term *ratio score* is used to denote s_i, which conceptually represents an adjustment factor based on how the client's representation rate compares to the global rate. The score s_i represents a multiplicative adjustment factor applied to the client's dataset size. Values $s_i > 1$ boost the client's influence (when it helps balance), while $s_i < 1$ reduce it (when it worsens imbalance). The score is clamped to $[0.5, 2.0]$ to prevent extreme weights. The clamp values $[0.5, 2.0]$ were chosen empirically to balance fairness improvement with stability:

   - Lower bound 0.5: Prevents completely excluding clients (minimum 50% of their base weight).
   - Upper bound 2.0: Prevents any single client from dominating (maximum 2× their base weight).

   These bounds ensure that ratio-based adjustments remain moderate, avoiding the instability that could arise from extreme weight ratios.

6. *Normalization* (lines 29-32). The raw client weights are normalized to sum to 1

7. *EMA Smoothing* (lines 34-40). To stabilize weights across rounds, FedCvg-Ratio applies Exponential Moving Average (EMA) to each client weight), based on a parameter λ:

$$w_i^r = \lambda \cdot w_i^{r-1} + (1 - \lambda) \cdot \hat{w}_i^r \tag{3.8}$$

where:

- $\lambda = 0$: only new weights are used (no memory).
- $\lambda = 0.5$: a balanced weight between old and new is computed.
- $\lambda = 1$: only previous weights are used (no update).

For new clients (first time observed), we set $w_i^r = \hat{w}_i^r$ (line 38).

8. *Aggregation* (line 42). The server aggregates client models using the EMA-smoothed weights.

3.5.3 FedCvg and FedCvg-Ratio Comparison

Table 3.3 summarizes the key differences between FedCvg and FedCvg-Ratio.

Table 3.3: Comparison of FedCvg and FedCvg-Ratio

Aspect	**FedCvg**	**FedCvg-Ratio**
Objective	Fixed coverage threshold	Global distribution balance
Weight Formula	$\exp(\alpha \cdot (n_i^{unpriv} - cov)) \cdot n_i$	$n_i \cdot (1 + \alpha \cdot norm\_diff)$
Reference	Static threshold *cov*	Dynamic rr^r_{global} (per-round)
Smoothing	None	EMA (Exponential Moving Average)
Adaptivity	Static	Adaptive to current distribution
Parameter Tuning	Requires dataset-specific *cov*	No threshold needed

The fundamental difference between the two approaches lies in their reference points for weight computation. FedCvg uses a static coverage threshold *cov* that must be specified before training and remains fixed throughout. This threshold represents an absolute number of samples from the protected group that clients should ideally possess. In contrast, FedCvg-Ratio uses a dynamic global ratio rr^r_{global} that is recomputed at each round based on the actual data distribution of participating clients. This makes FedCvg-Ratio inherently adaptive: as the composition of participating clients changes (especially with Parity Sampling), the reference point adjusts automatically.

The weight formulas reflect this difference. FedCvg's exponential formula amplifies or dampens client influence based on the distance from the fixed threshold *cov*. FedCvg-Ratio's formula instead adjusts weights based on how much a client's local ratio deviates from the current global average, using a normalized difference to ensure consistent effects

across different distributions. Additionally, FedCvg-Ratio incorporates EMA smoothing to prevent abrupt weight changes between rounds, which is particularly important when the global ratio fluctuates due to varying client participation.

From a practical standpoint, FedCvg requires careful tuning of the *cov* parameter for each dataset, as an inappropriate threshold can harm both fairness and accuracy. FedCvg-Ratio eliminates this burden by automatically adapting to the data distribution, making it more robust and easier to deploy across different scenarios.

FedCvg-Ratio offers several advantages over the original FedCvg:

1. *No threshold specification*: Eliminates the need to specify a coverage threshold *cov*, which may not generalize across datasets or training stages.
2. *Dynamic adaptation*: Automatically adapts to the current data distribution as it evolves during training. The reference point r_{global} is recomputed each round based on participating clients.
3. *Stability*: EMA smoothing prevents abrupt weight changes that could destabilize training, especially important when combined with Parity Sampling.
4. *Robustness*: The ratio-based approach is inherently robust to varying dataset sizes and distributions. A client with 60% unprivileged samples receives similar relative treatment whether the global average is 30% or 40%.

3.6 Parity Sampling

Parity Sampling is a client selection strategy that can be combined with any federated learning algorithm to improve representation balance. Introduced in Brocchi's thesis [Bro23], it strategically selects which clients participate in each training round based on their demographic composition.

3.6.1 Introduction

While the algorithms discussed so far focus on changing weights to influence the impact of each client in the final aggregation, Parity Sampling addresses a complementary question: which clients should participate in each round? The key insight is that when only a fraction of clients can participate per round (i.e., fraction fit lower than 1 and $|S_r| < C$), strategic selection can amplify fairness improvements by preferentially choosing clients that

help balance the global distribution. It thus corresponds to a specific implementation of the ClientSelection() function introduced in the previously presented algorithms.

Parity Sampling works by tracking the global distribution of protected groups across all selected clients. At each round (after the first), it identifies which group is currently under-represented and preferentially selects clients that have more samples from that group. This creates a feedback loop: as the global distribution becomes more balanced, the selection pressure decreases, eventually converging to random selection when perfect balance is achieved.

3.6.2 Description

The Parity Sampling algorithm proceeds as shown in Algorithm 6.

The approach relies on the probability p of using Parity Sampling instead of random sampling in any round. For example, $p = 0.5$ means 50% chance of Parity Sampling and 50% chance of random selection. Besides p, similarly to the ClientSampling() function, it also takes the fraction fit ff, the set of clients C as parameters, together with round r at which is it called.

We also assume the server maintains a global client registry to track dataset statistics (n_i, n_i^{unpriv}) for all observed clients across rounds. The registry is initially empty and is populated dynamically as clients are selected and observed during training. When a client i participates in a round for the first time, its statistics are added to the registry and used in subsequent rounds for scoring.

We now explain each step of the algorithm in detail, using the notation introduced in Table 3.1.

1. *Round 1: random sampling* (lines 2-3). In the first round, clients are always randomly selected because the client registry is still empty and no distribution information is available yet (line 3). This initial random selection establishes baseline statistics that will populate the registry for subsequent rounds.

2. *Subsequent rounds: probabilistic selection* (lines 4-21). From round 2 onward, with probability $(1 - p)$, the algorithm uses random sampling (line 5); with probability p, it applies Parity Sampling (lines 7-20). This probabilistic approach balances fairness improvement with model diversity.

3. *Distribution tracking* (lines 7-8). The server computes the global distribution from all previously observed clients using the global client registry (not just the ones selected in the current round). In the formulas, *observed* corresponds to the set of clients

Algorithm 6 PARITYSAMPLING(p, ff, C, r)

1: **Server:**
2: **if** $r == 1$ **then**
3: $S_r \leftarrow$ RandomClientSampling(ff, C)
4: **else if** random() $\geq p$ **then**
5: $S_r \leftarrow$ RandomClientSampling(ff, C)
6: **else**
7: $n^{unpriv} \leftarrow \sum_{i \in observed} n_i^{unpriv}$
8: $n^{priv} \leftarrow \sum_{i \in observed} (n_i - n_i^{unpriv})$
9: **if** $n^{unpriv} == n^{priv}$ **then**
10: $S_r \leftarrow$ RandomClientSampling(ff, C)
11: **else**
12: **for** each client $i \in C$ **do**
13: **if** $n^{unpriv} > n^{priv}$ **then**
14: $score_i \leftarrow n_i^{priv}$
15: **else**
16: $score_i \leftarrow n_i^{unpriv}$
17: **end if**
18: **end for**
19: Sort clients by $score_i$ descending
20: $S_r \leftarrow$ top $(ff \cdot |C|)$ clients
21: **end if**
22: **end if**
23: Update client registry with data of selected clients
24: return S_r

selected up to the current round (i.e., those for which the corresponding position in the client registry has been updated).

The registry is necessary for PARITY SAMPLING because the algorithm needs to track the cumulative distribution across all clients seen so far, not just those in the current round. This allows it to make informed selection decisions based on the global imbalance.

4. *Balanced distribution check* (lines 9-10). If the distribution is perfectly balanced ($n^{unpriv} == n^{priv}$) (line 9), the algorithm falls back to random sampling (line 10).

5. *Client scoring* (lines 11-21). If the distribution is not perfectly balanced, each client $i \in C$ is assigned a score based on which group is under-represented (lines 13-17):

   The intuition is:

- If group 0 is over-represented ($n^{unpriv} > n^{priv}$): select clients with more privileged samples to balance (and set $score_i$ to n_i^{priv}).
- If group 1 is over-represented ($n^{unpriv} < n^{priv}$): select clients with more unprivileged samples to balance (and set $score_i$ to n_i^{unpriv}).

6. *Client selection* (lines 19-20). Clients are sorted by score in descending order, and the top $ff \cdot |C|$ clients are selected to form the subset S_r. This ensures that clients most helpful for balancing the distribution are prioritized.

7. *Final operations* (lines 23-24). The client registry is updated and S_r is returned.

3.6.3 Limitations

Parity Sampling can significantly affect which clients participate in training, thus having an impact on selection fairness (see Section 2.1):

- *Clients never selected*: Clients with very poor representation of both groups (e.g., only 10 samples total) will consistently receive low scores and may never be selected. This is intentional: such clients contribute little to balancing the distribution.

- *Clients always selected*: Clients with excellent representation of the currently under-represented group will consistently receive high scores. For example, if group 0 is globally under-represented at 30%, a client with 80% group 0 samples will be prioritized in most rounds.

- *Dynamic selection*: As the global distribution becomes more balanced through training, the selection pressure decreases. A client that was highly prioritized early (when imbalance was severe) may become less prioritized later (when balance improves).

- *Partial client participation*: Parity Sampling is only effective when not all clients participate in every round (i.e., $ff! = 1$ and therefore $|S_r| < |C|$); this allows the algorithm to prioritize clients with better representation of the under-represented group. On the other hand, if all clients are selected in each round ($|S_r| = |C|$), Parity Sampling has no practical effect because there is no selection to perform—all clients participate regardless of their scores. The scoring mechanism becomes irrelevant when there is no subset to choose.

Additional limitations are the following:

- *Reduced diversity*: By consistently favoring certain clients, Parity Sampling may reduce model diversity and potentially harm convergence on non-IID data.

- *Partial participation*: The approach is only effective when not all clients participate in every round ($|S_r| < |C|$), which may slow convergence compared to full participation.
- *Cold start*: New clients that haven't been observed yet cannot be scored, potentially delaying their participation.

3.7 Local Debiasing

Local debiasing is a client-side fairness intervention that can be combined with any of the server-side algorithms (FedAvg, FairFed, FedCvg, FedCvg-Ratio). Unlike the server-side approaches that modify aggregation weights, local debiasing is a pre-processing approach to limit algorithmic bias (see Section 1.2.2); it operates during local training by reweighting samples based on their demographic group membership.

3.7.1 Introduction

Local debiasing addresses a fundamental challenge in federated learning: clients may have highly skewed distributions of (A, Y) combinations, where A is the sensitive attribute and Y is the label. For example, a client might have many samples with $(A = 1, Y = 1)$ (privileged group, positive label) but few with $(A = 0, Y = 0)$ (unprivileged group, negative label). This imbalance can lead to biased local models that, when aggregated, produce a biased global model.

The key insight behind local debiasing is that we can correct for these imbalances *during training* by assigning different weights to training samples based on their group membership. Samples from underrepresented (A, Y) cells receive higher weights, while samples from overrepresented cells receive lower weights. This effectively balances the contribution of each demographic group during local training, even when the underlying data distribution is skewed.

The approach is particularly attractive in federated settings because it operates independently at each client without requiring coordination with the server or other clients. This preserves the privacy guarantees of federated learning while still addressing representation imbalances. Each client can apply local debiasing based solely on its own data distribution, without revealing demographic information to the server.

3.7.2 Description

We now describe the reweighting approach proposed by Kamiran and Calders [KC12], a pre-processing technique introduced in Section 1.2.2 that modifies the training data distribution to achieve statistical parity.

For clarity, recall that $A \in \{0, 1\}$ denotes the binary sensitive attribute (0 = unprivileged, 1 = privileged) and $Y \in \{0, 1\}$ denotes the binary label, as introduced in Chapter 1. We denote by $w(A, Y)$ the weight assigned to samples in cell (A, Y).

Local debiasing operates in several steps, described in the following.

1. *Objective.* The goal is to assign weights such that the weighted joint distribution $P_{weighted}(A, Y)$ becomes independent, i.e., the sensitive attribute and label are statistically independent in the weighted distribution:

$$P_{weighted}(A, Y) = P(A) \cdot P(Y) \tag{3.9}$$

   This ensures that the model trained on the reweighted data does not learn spurious correlations between the sensitive attribute and the label.

2. *Weight Formula.* The weight for each sample with sensitive attribute value A and label Y is computed as:

$$w(A, Y) = \frac{P(Y) \cdot P(A)}{P(A, Y)} \tag{3.10}$$

   where:

   - $P(Y)$ is the marginal probability of label Y in the client's dataset
   - $P(A)$ is the marginal probability of sensitive attribute A in the client's dataset
   - $P(A, Y)$ is the joint probability of (A, Y) in the client's dataset

   The intuition is straightforward: if a cell (A, Y) is overrepresented (high $P(A, Y)$), its samples receive lower weights. If a cell is underrepresented (low $P(A, Y)$), its samples receive higher weights. The formula ensures that after reweighting, the expected proportion of each cell matches the product of the marginals.

3. *Implementation.* During local training, each client:

   (a) Computes the empirical probabilities $P(A)$, $P(Y)$, and $P(A, Y)$ from its local dataset

Table 3.4: Local debiasing weight computation example

Cell (A, Y)	Count	P(A,Y)	P(A)	P(Y)	Weight
(0, 0)	10	0.10	0.30	0.40	$0.30 \times 0.40/0.10 = \mathbf{1.20}$
(0, 1)	20	0.20	0.30	0.60	$0.30 \times 0.60/0.20 = \mathbf{0.90}$
(1, 0)	30	0.30	0.70	0.40	$0.70 \times 0.40/0.30 = \mathbf{0.93}$
(1, 1)	40	0.40	0.70	0.60	$0.70 \times 0.60/0.40 = \mathbf{1.05}$

(b) Calculates weights $w(A, Y)$ for each of the four cells: $(0, 0)$, $(0, 1)$, $(1, 0)$, $(1, 1)$

(c) Applies these weights during training by passing them to the loss function (e.g., weighted cross-entropy)

Example 3.2 *Consider a client with 100 samples distributed as described in Table 3.4. The underrepresented cell* $(A = 0, Y = 0)$ *receives the highest weight (1.20), effectively increasing its influence during training. The overrepresented cell* $(A = 1, Y = 1)$ *receives a weight close to 1.0. After reweighting, the effective distribution becomes more balanced.*

4. *Integration with server-side algorithms.* Local debiasing is orthogonal to the choice of server-side aggregation algorithm. It can be enabled or disabled independently for any algorithm (FedAvg, FairFed, FedCvg, FedCvg-Ratio). This modularity allows us to study the complementary effects of client-side and server-side fairness interventions.

3.7.3 Limitations

While local debiasing is a powerful technique, it has several limitations:

- *Need for a sufficient sample set per cell*: The reweighting approach requires that each client has at least some samples in each (A, Y) cell. If a cell is completely empty, the weight cannot be computed. In practice, this can be problematic with highly skewed data distributions.
- *May not address all fairness metrics*: Local debiasing aims for statistical parity (independence between A and Y), but this may not be sufficient for other fairness notions like equal opportunity or equalized odds. The technique does not directly optimize for these metrics.

- *Potential for overfitting*: Aggressive reweighting can lead to overfitting, especially when underrepresented cells have very few samples. The high weights assigned to these samples can cause the model to overfit to their specific characteristics.
- *No Coordination across clients*: Each client applies local debiasing independently based on its own data distribution. This lack of coordination means that clients may apply conflicting reweighting schemes, potentially reducing the effectiveness of the technique at the global level.
- *Interaction with server-side methods*: The interaction between local debiasing and server-side fairness interventions is not always additive. In some cases, combining both approaches may not yield better results than using either alone, as we will show in our experimental evaluation.

Despite these limitations, local debiasing remains a valuable tool in the fairness-aware federated learning toolkit, particularly when combined judiciously with server-side approaches.

Chapter 4

Experimental Setup

This chapter presents the experimental framework designed to evaluate the fairness-aware federated learning algorithms described in Chapter 3. We begin by introducing the Flower framework that provides the foundation for our implementation (Section 4.1), then describe the datasets (Section 4.2) and data partitioning strategies (Section 4.3) that shape the heterogeneity conditions under which we evaluate the algorithms. Section 4.4 presents the algorithm variants we compare, Section 4.5 outlines the experimental design and parameter configurations, Section 4.7 describes the experimental scenarios we investigate Section 4.6 summarizes the metrics used for the comparison and Section 4.8 provides information for reproducing the obtained results.

4.1 The Flower Framework

Implementing and evaluating federated learning algorithms requires a robust framework that can handle the complexities of distributed training while remaining flexible enough to accommodate custom aggregation strategies. For this thesis, we selected Flower [BTM+20], an open-source federated learning framework that has gained significant traction in both research and industry settings.

The choice of Flower over alternatives such as TensorFlow Federated (TFF) [BEG+19], PySyft [RTD+18], or FedML [HLS+20] was driven by several considerations. First, Flower is *framework-agnostic*: it works seamlessly with PyTorch [PGM+19], TensorFlow [ABC+16], JAX [BFH+18], and other machine learning libraries, allowing us to leverage PyTorch's flexibility for model development.

Second, Flower provides a clean separation between the federated learning logic and the underlying infrastructure, making it straightforward to implement custom aggregation

strategies—a crucial requirement for fairness-aware algorithms like FairFed and FedCvg. Third, Flower supports both simulation model (where all clients run in a single process) and distributed deployment, enabling us to develop and test locally before scaling to larger experiments.

Perhaps most importantly for our work, Flower's strategy abstraction provides a natural extension point for implementing fairness-aware aggregation. The *strategy abstraction* in Flower is a design pattern that encapsulates the server-side logic for client selection and model aggregation. By implementing a custom Strategy class, we can override specific methods (e.g., `aggregate_fit()`) to define fairness-aware weighting schemes while inheriting robust default behavior for client communication, model serialization, and failure handling. This approach allows us to implement custom aggregation strategies without modifying core framework code.

4.1.1 Architecture and Training Flow

Flower implements the standard federated learning architecture with a central server coordinating multiple distributed clients. Figure 4.1 illustrates the key components and their interactions during a typical training round.

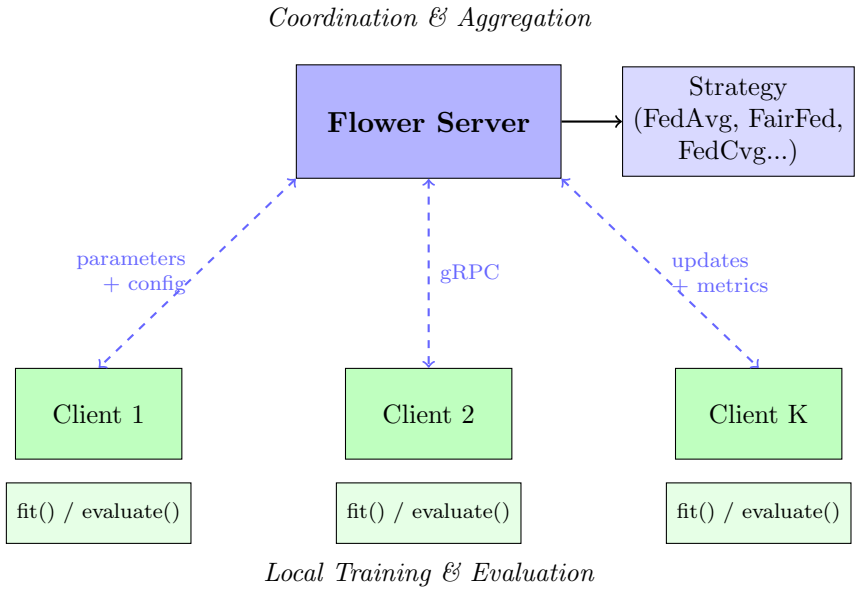


Figure 4.1: Flower framework architecture

The server orchestrates the entire federated learning process through a *Strategy* object that defines how clients are selected and how their updates are aggregated. For fairness-aware algorithms, we override the aggregation method to implement custom weighting

schemes: according to what presented in Chapter 3, FairFed computes weights based on fairness gaps, FedCvg uses coverage-based weights, and FedCvg-Ratio applies ratio-based adjustments with EMA smoothing.

Each client wraps a local machine learning model and dataset, implementing methods for training (`fit()`) and evaluation (`evaluate()`). For fairness-aware algorithms, clients extend these methods to return additional metadata: FairFed clients report fairness metrics computed on the global model before training, while FedCvg and FedCvg-Ratio clients report group counts that the server uses for representation-based weighting (see also Table 2.1).

Our implementation follows a modular structure where each algorithm is implemented as a separate Flower application. Each application consists of a `task.py` file containing dataset loading and model definitions, a `client_app.py` file implementing the client logic, and a `server_app.py` file defining the custom aggregation strategy. This organization ensures consistency across algorithms while keeping algorithm-specific logic cleanly separated.

4.2 Datasets

We evaluate our algorithms on two well-established fairness-sensitive datasets commonly used in the fairness literature: Adult Income [KB96] and COMPAS Recidivism [Pro16].

Both datasets involve *binary classification tasks* where the goal is to predict a binary outcome (income level or recidivism risk) while ensuring fairness with respect to sensitive attributes. The *Adult Income* dataset is derived from the 1994 U.S. Census database, with the task of predicting whether an individual's annual income exceeds \$50,000 (binary classification: $\leq$\$50K vs. >\$50K). The raw dataset contains 48,842 samples with 14 features (6 continuous, 8 categorical). After preprocessing—which includes handling missing values through mode imputation, one-hot encoding categorical features, and removing the census weight feature (`fnlwgt`)—we obtain 48,842 samples with approximately 108 features. The class distribution is imbalanced (75% $\leq$\$50K, 25% >\$50K). Attribute *sex* is used as sensitive attribute, with females corresponding to 33% of the dataset and males to 67%.

The *COMPAS Recidivism* dataset contains criminal history data used to predict two-year recidivism risk (binary classification: recidivism vs. no recidivism). This dataset gained attention following ProPublica's 2016 investigation [ALMK16] revealing racial bias in the COMPAS algorithm. The raw dataset contains 7,214 samples. After applying the ProPublica screening filter ($\pm$30 days between arrest and COMPAS screening) and standard preprocessing, we obtain 6,172 samples with 11 features, of which approximately 5,000 males (81%) and 1,172 females (19%). The class distribution is more balanced (54% no recidivism, 46% recidivism), but the sensitive attribute distribution is highly skewed,

with males designated as the unprivileged group and females as the privileged group. Also in COMPAS, *sex* is used as sensitive attribute.

The designation of privileged and unprivileged groups in the two datasets follows the definition of the favorable outcome within each task.

For the Adult dataset, the favorable outcome is defined as earning an income greater than \$50K. Empirical evidence shows that women are significantly underrepresented in the high-income category. Therefore, females are usually defined as the unprivileged group in this context, as they are less likely to receive the favorable outcome.

For the COMPAS dataset, the favorable outcome is defined as being classified as low risk. Since male defendants receive high-risk classifications more frequently than female defendants, males are sometimes treated as the unprivileged group with respect to this outcome. This allows fairness metrics to capture potential disparities in the probability of receiving a favorable prediction.

It is important to emphasize that the terms "privileged" and "unprivileged" are used in a technical sense required for the computation of group fairness metrics. These labels do not imply normative judgments, but rather identify the group that is less likely to receive the favorable outcome in each specific prediction task.

These two datasets provide complementary characteristics for our evaluation. Adult offers a larger sample size with moderate sensitive attribute imbalance but significant class imbalance, while COMPAS provides a smaller dataset with balanced classes but extreme sensitive attribute imbalance. Additionally, Adult females are the minority group while in COMPAS males are the majority ones. This diversity allows us to evaluate algorithm robustness across different data characteristics.

4.3 Data Partitioning Strategies

A critical aspect of Federated Learning experiments is how data is distributed across clients, as the partitioning method directly shapes the heterogeneity of local datasets and the fairness challenges that emerge during training. We employ two families of partitioning strategies, each creating different types of non-IID distributions.

4.3.1 Dirichlet Partitioning

The Dirichlet distribution has become the de facto standard for creating non-IID partitions in Federated Learning research [HQB19]. Following the experimental setup from the FairFed paper [EYH+23], we partition data based solely on the sensitive attribute. For

each sensitive attribute value $a \in \{0, 1\}$, we independently sample proportions from a symmetric Dirichlet distribution with concentration parameter α:

$$p^{(a)} \sim \text{Dir}(\alpha, \ldots, \alpha) \tag{4.1}$$

The concentration parameter α controls the degree of heterogeneity. We test five values ranging from highly non-IID ($\alpha = 0.1$) to effectively IID ($\alpha = 5000$), with intermediate values at $\alpha \in \{0.2, 0.5, 10\}$. To prevent empty clients with low α values, we enforce a minimum of 100 samples per client.

4.3.2 Coverage-Based Partitioning

The coverage-based partitioning methods were introduced in the Master's thesis of Martina Brocchi [Bro23] as experimental setups to evaluate the FedCvg algorithm under controlled heterogeneity conditions. Unlike Dirichlet partitioning where heterogeneity emerges stochastically, these methods create controlled non-IID distributions based on an explicit coverage constraint, allowing precise control over which clients have good representation of the protected group.

The coverage constraint (see Section 3.4 for a definition) defines the minimum number of unprivileged samples a client must have to be considered a "good client". Given a coverage threshold *cov* and tolerance τ, the minimum samples threshold is $c \cdot (1 - \tau)$. For example, with $cov = 1000$ and $\tau = 0.1$, a good client must have at least $1000 \cdot 0.9 = 900$ unprivileged samples.

The number of clients is computed automatically based on the dataset size and coverage constraint. Let n^{unpriv} be the total number of unprivileged samples in the dataset, and let p_{good} be the desired percentage of good clients (e.g., 0.5 for 50%). The algorithm first determines how many good clients can be created: $k^{good} = \lfloor n^{unpriv}/cov \rfloor$. Then, the total number of clients is $k = k^{good}/p^{good}$. The remaining $k - k^{good}$ clients are designated as "bad clients" with poor representation of the unprivileged group.

We use two variants, described in details in [Bro23]: **cov_same_size**, where all clients receive approximately the same number of samples but vary in their group distribution, and **cov_diff_size**, which allows both client sizes and group distributions to vary, simulating a more realistic scenario.

In our experiments, we test three coverage values for each dataset, chosen to create different levels of representation challenge:

- **Adult dataset** ($n^{unpriv} = 16,117$ females): $cov \in \{4497, 2570, 1999\}$

  - $cov = 4497$ (27.9% of unprivileged samples): high coverage (few clients, each with many unprivileged samples).
  - $cov = 2570$ (15.9% of unprivileged samples): medium coverage (moderate number of clients).
  - $cov = 1999$ (12.4% of unprivileged samples): low coverage (more clients, each with fewer unprivileged samples).

- **COMPAS dataset** ($n^{unpriv} = 4,998$ males): $cov \in \{1067, 610, 474\}$
  - $cov = 1067$ (21.3% of unprivileged samples): high coverage.
  - $cov = 610$ (12.2% of unprivileged samples): medium coverage.
  - $cov = 474$ (9.5% of unprivileged samples): low coverage.

These values are proportional to the dataset sizes, with Adult coverage values being approximately 4.21 of COMPAS values (e.g., $4497/1067 \approx 4.21$). This ratio is chosen to create comparable representation challenges across datasets, accounting for their different sizes and distributions. Lower coverage values create more clients with less representation of the protected group, making the fairness challenge more difficult.

For all coverage-based experiments, we use $\tau = 0.1$ and $p_{good} = 0.5$ (50% good clients), meaning half the clients have adequate representation of the unprivileged group while the other half have poor representation. This creates a balanced scenario where the aggregation algorithm must decide how much to weight clients with good or poor coverage. We also set tolerance $\tau = 0.1$ (10%), allowing a small margin below the exact coverage threshold.

The choice of 50% good clients represents a middle ground: with fewer good clients (e.g., 25%), the fairness challenge becomes more severe as most clients have poor representation; with more good clients (e.g., 75%), the challenge becomes easier. While varying this percentage could provide additional insights into algorithm behavior under different representation scenarios, we focus on the 50% case as it represents a realistic and challenging setting. Exploring different percentages of good clients (25%, 50%, 75%) remains an interesting direction for future work.

4.4 Compared Algorithms

Our experimental evaluation compares the three main algorithms presented in Chapter 3 (FedAvg, FairFed, FedCvg, and FedCvg-Ratio). For each algorithm, we test multiple configurations by varying key parameters and combining each base technique with two orthogonal techniques, namely Parity Sampling (PS in the following) under specific

probabilities and fraction fits (see Section 3.6) and Local Debiasing (LD in the following) considering the Kamiran and Calders reweighting approach (see Section 3.7).

When considering parity sampling, we assume it is used in all the rounds of a reference approach (thus parameter r is not specified). In addition, we consider one single value for the fraction fit ff equal to 0.7 will be considered; for this reason, we will not highlight this parameter in the name of the variants. This value represents a balance between strategic client selection (allowing the server to choose which 70% of clients participate) and maintaining sufficient client participation for stable training. Values closer to 1.0 would limit the server's ability to strategically select clients, while values significantly lower than 0.7 would reduce client participation too much, potentially harming convergence. When parity sampling is not used, all clients are considered at each round (i.e., ff is equal to 1). Finally, we do not list parameter C, which depends on the chosen data partitioning and set-up.

More precisely, we consider the following approaches:

- FedAvg *variants*:
  - FedAvg: baseline without local debiasing and without parity sampling.
  - FedAvg+LD: variant with local debiasing.
  - FedAvg+PS(p): variant with parity sampling with probability $p \in [0, 1]$, where $p = 0$ means random selection and $p = 1$ means always selecting clients with more unprivileged samples.
  - FedAvg+LD+PS(p): variant with local debiasing and parity sampling with probability p.
- FairFed *variants*:
  - FairFed(ϕ): variant considering metric ϕ ($\phi \in \{SPD, EOD, ACC\_DIFF\}$) without local debiasing (differently from what proposed in [EYH+23]) and without parity sampling.
  - FairFed(ϕ)+LD: variant with local debiasing.
  - FairFed(ϕ)+PS(p): variant with parity sampling with probability p.
  - FairFed(ϕ)+LD+PS(p): variant with local debiasing and parity sampling with probability p.

  For the FairFed execution, a value for parameter β, controlling the weight update rate (see Section 3.3) has to be set. As recommended in the original paper, we set β to 1 and we do not specify it in the algorithm notation.
- FedCvg *variants*:

  - FedCvg(α_{cov}): variant for a certain value for α (representing the coverage sensitivity parameter, controlling how much coverage affects weights).
  - FedCvg(α_{cov})+LD: variant with local debiasing.
  - FedCvg(α_{cov})+PS(p): variant with parity sampling with probability p.
  - FedCvg(α_{cov})+LD+PS(p): variant with local debiasing and parity sampling with probability p.

- FedCvg-Ratio *variants*:
  - FedCvg-Ratio(α_{rr},λ): variant for a certain value for α (representing the ratio sensitivity parameter, controlling how much the ratio affects weights) and λ (representing the EMA smoothing factor).
  - FedCvg-Ratio(α,λ)+LD: variant with local debiasing.
  - FedCvg-Ratio(α_{rr},λ)+PS(p): variant with parity sampling with probability p.
  - FedCvg-Ratio(α_{rr},λ)+LD+PS(p): variant with local debiasing and parity sampling with probability p.

  Based on some preliminarly performed experiments, we choose 0.5 as value for α_{rr} and do not further specify it in the algorithm notation.

In the following, we refer to FedCvg and FedCvg-Ratio variants as to *representation-based* or *representation-aware* approaches.

4.5 Experimental Design

Our experimental evaluation is organized around a systematic exploration of algorithm behavior under different configurations. Experiments run on standard hardware using CPU-bound computation with 8GB+ RAM. The Flower framework handles client simulation in a single process, enabling reproducible experiments without distributed infrastructure. All experiment configurations and results are saved in a structured format that facilitates systematic analysis and comparison.

All experiments share a common base configuration designed to ensure fair comparison across algorithms. Table 4.1 summarizes the key parameters:

- *Training parameters:* We train for 100 rounds with 1 local epoch per round to prevent excessive client drift while allowing sufficient training time. We test three learning

Table 4.1: Base experimental configuration

Parameter	**Value**
Training parameters:	
Number of rounds	100
Local epochs per round	1
Learning rate	$\{0.1, 0.01, 0.001\}$
Batch size	32
Optimizer	SGD
Model:	
Architecture	Logistic Regression
Data split:	
Train/Test split	80% / 20%
Client configuration:	
Number of clients (Dirichlet)	5
Number of clients (Coverage)	Computed automatically based on coverage value
Reproducibility:	
Random seeds	$\{42, 123, 456, 789, 101112\}$

rates ($\{0.1, 0.01, 0.001\}$) to find the optimal value for each dataset and algorithm combination. The batch size is fixed at 32, and we use standard stochastic gradient descent (SGD) as the optimizer.

- *Model:* We use logistic regression as our model architecture, chosen for its interpretability, fast training, and widespread use in fairness research. This simple model allows us to focus on the effects of different aggregation strategies without confounding factors from complex model architectures.

- *Data split:* We use a standard 80/20 train/test split. The training data is distributed across clients according to the partitioning strategy, while the test set remains centralized at the server for consistent evaluation.

- *Client configuration:* The number of clients differs between partitioning strategies. For Dirichlet partitioning, we use a fixed number of 5 clients, following the experimental setup from the FairFed paper [EYH+23]. For coverage-based partitioning, the number of clients is computed automatically based on the coverage value and percentage of good clients (see Section 4.3), typically resulting in 7-15 clients depending on the coverage value.

- *Reproducibility:* Each configuration is run with five different random seeds ($\{42, 123, 456, 789, 101112\}$), allowing us to report mean and standard deviation for all metrics and distinguish genuine improvements from random fluctuations.

Table 4.2: Parameter values

Algorithm	**Parameter**	**Values**
FairFed	β	1
	ϕ	$\{SPD, EOD, ACC\_DIFF\}$
FedCvg	α_{cov}	$\{0.0001, 0.01\}$
	cov	dataset based (see Section 4.3.2)
FedCvg-Ratio	α_{rr}	0.5
	λ	$\{0.5, 0.9\}$
PS	p	$\{0.5, 0.7\}$
	ff	0.7

Table 4.2 summarizes the parameters used to define the algorithm variants. As previously discussed, we consider the β value proposed in [EYH+23], value 0.5 for α_{rr}, and value 0.7 for the fraction fit under parity sampling. We test two EMA smoothing factors, with $\lambda \in \{0.5, 0.9\}$, where $\lambda = 0.5$ provides moderate smoothing and $\lambda = 0.9$ provides strong smoothing to prevent abrupt weight changes.

We remark that, within the FedCvg family, parameter α_{cov} controls the strength of the coverage-based correction: the high the value the stronger the coverage-based impact. On the other hand, the coverage threshold *cov* is differently set based on the partitioning method. Under coverage-based partitioning, *cov* coincide with the threshold of a coverage constraint on the unprivileged group, defined in a domain-dependent way (e.g., 4497, 2570, or 1999 for Adult). Under Dirichlet partitioning, *cov* is set to the mean unprivileged sample count per client, computed by: (1) performing the Dirichlet partition, (2) counting unprivileged samples in each client, and (3) taking the mean of these counts. This adaptive approach allows FedCvg to apply coverage-based weighting consistently across different partitioning strategies.

4.6 Evaluation Metrics

We evaluate algorithms using both performance and fairness metrics. Performance is measured through accuracy, loss, and precision. Fairness is assessed using the metrics defined in Chapter 1: Statistical Parity Difference (SPD), Equal Opportunity Difference (EOD), Average Odds Difference (AOD), and Accuracy difference (ACC_DIFF). All fairness metrics have values closer to 0 indicating better fairness. All metrics are computed on a held-out global test set to ensure unbiased evaluation.

In addition, to evaluate the overall trade-off between accuracy and fairness, we introduce a new combined metric that captures both objectives simultaneously. The metric, called *Fairness-Accuracy Score (FAS)* is defined as follows:

$$\text{FAS} = \text{Accuracy} \times \left(1 - \frac{\text{EOD} + \text{SPD} + \text{AOD} + \text{AccDiff}}{4}\right) \tag{4.2}$$

This metric rewards methods that achieve both high accuracy and low fairness violations across all four fairness metrics. Higher FAS values indicate better overall performance, balancing both accuracy and fairness. More precisely, the FAS metric has the following properties:

- When all fairness metrics, including accuracy difference, are 0 (perfect fairness), FAS equals the accuracy. On the other hand, when the average of fairness metrics is 1 (maximum unfairness), FAS equals 0.
- When accuracy is 0 (minimum accuracy), FAS equals 0; when it is 1 (maximum accuracy), FAS equals the 1 minus the average fairness metric value.

4.7 Experimental Scenarios

Our evaluation consists of two main experiments, described in the following.

Experiment 1 – Impact of dataset heterogeneity. The goal of the first experiment is to compare the reference algorithms with respect to different levels of client dataset heterogeneity, defined in terms of different Dirichlet partitioning, as described in Section 4.3.1.

The considered Dirichlet partitionings span from highly non-IID distributions ($\alpha = 0.1$, where clients have very skewed sensitive attribute distributions) to effectively IID distributions ($\alpha = 5000$, where all clients have similar sensitive attribute distributions). For all the algorithm versions, we consider all the parameter values pointed out in Table 4.2 and $\alpha \in \{0.1, 0.2, 0.5, 10, 5000\}$, on both Adult and COMPAS datasets.

We analyse the behaviour of the considered techniques to determine: (a) their trend for each considered metric and different heterogeneity levels; (b) the impact of local debiasing, determining whether client-side and server-side interventions are complementary or redundant; (c) the impact of parity sampling; (d) the interaction between local debiasing and parity sampling; (e) their behaviour with respect to the combined performance metric FAS.

Experiment 2 – Impact of coverage-based partitioning. The goal of the second experiment is to compare the reference algorithms with respect to different levels of client dataset heterogeneity, defined in terms of coverage-constraints, as described in Section 4.3.2.

To this aim, we consider the two partitioning variants presented in Section 4.3.2 (cov_same_size, cov_diff_size) and the coverage thresholds discussed in Section 4.3.2 (Adult: $\{4497, 2570, 1999\}$; COMPAS: $\{1067, 610, 474\}$). For all the algorithm versions, we consider all the parameter values pointed out in Table 4.2, on both Adult and COMPAS datasets.

The analysis is performed as explained for Experiment 1.

The analysis also identifies the impact of the parameter values (α_{cov} and λ) on FedCvg and FedCvg-Ratio on the considered metrics.

4.8 Reproducibility

To ensure full reproducibility of our experimental results, we provide a comprehensive open-source implementation available on GitHub.[1] The complete codebase includes implementations of all algorithms (FedAvg, FairFed, FedCvg, FedCvg-Ratio) as Flower applications, data partitioning utilities for both Dirichlet and coverage-based methods, experiment orchestration scripts, and analysis tools. The Adult and COMPAS datasets are automatically downloaded from their public sources (UCI Machine Learning Repository [KB96] and ProPublica [Pro16]) during the first run, with all preprocessing steps documented in the code.

The implementation requires Python 3.9+ and uses the Flower framework [BTM+20] for federated learning simulation. After cloning the repository and installing dependencies via `pip install -r requirements.txt`, all experiments can be reproduced using the `run_grid.py` script, which orchestrates multiple runs with different configurations and automatically saves results in a structured format. For example, to reproduce Experiment 1 with Dirichlet partitioning:

```
python3 src/experiments/run_grid.py \
    --run fedavg fairfed fedcvg fedcvg-ratio \
    --dataset adult compas \
    --partition-methods dirichlet \
    --dirichlet-alphas 0.1 0.2 0.5 10 5000 \
    --seeds 42 123 456 789 101112 \
    --rounds 100
```

Coverage-based experiments can be reproduced by specifying `--partition-methods cov_same_size cov_diff_size` along with the appropriate coverage values (e.g., `--coverages 4497 2570 1999` for Adult dataset).

[1] https://github.com/leonardogonfiantini/federating_learning_master_thesis

4.8.1 Algorithm-Specific Parameters

Each algorithm accepts specific command-line flags that control its behavior, including options for local debiasing (Section 3.7) and parity sampling (Section 3.6):

- FedAvg family:
  - `--fedavg-local-debias true/false`: Enable Kamiran-Calders reweighting (Section 3.7)
  - `--fedavg-parity-ps 0.5 0.7`: Parity Sampling probability (Section 3.6)
  - `--fedavg-fraction-fits 0.7`: Client selection fraction (required with Parity Sampling)
- FairFed family:
  - `--fairfed-betas 1.0`: Weight update rate (β parameter, Section 3.3)
  - `--fairfed-metrics eod spd acc_diff`: Fairness metric to optimize
  - `--fairfed-local-debias true/false`: Enable local debiasing
  - `--fairfed-parity-ps 0.5 0.7`: Parity Sampling probability
  - `--fairfed-fraction-fits 0.7`: Client selection fraction
- FedCvg family:
  - `--fedcvg-coverage-alphas 0.0001 0.01`: Coverage sensitivity parameter (α_{cov}, Section 3.4)
  - `--fedcvg-local-debias true/false`: Enable local debiasing
  - `--fedcvg-parity-ps 0.5 0.7`: Parity Sampling probability
  - `--fedcvg-fraction-fits 0.7`: Client selection fraction
- FedCvg-Ratio family:
  - `--fedcvg-ratio-coverage-alphas 0.5`: Ratio sensitivity parameter (α_{rr}, Section 3.5)
  - `--fedcvg-ratio-ema-lambdas 0.5 0.9`: EMA smoothing factor (λ, Section 3.5)
  - `--fedcvg-ratio-local-debias true/false`: Enable local debiasing
  - `--fedcvg-ratio-parity-ps 0.5 0.7`: Parity Sampling probability
  - `--fedcvg-ratio-fraction-fits 0.7`: Client selection fraction

The script automatically generates all combinations of these parameters and runs each configuration with all specified random seeds, ensuring comprehensive coverage of the experimental space.

4.8.2 Output Structure

Results are saved in timestamped directories under `experiments/` with a hierarchical structure that organizes data by dataset, algorithm, and configuration. Each experiment run creates the following directory tree:

```
experiments/YYYYMMDD_HHMMSS/
  experiment_config.json       # Complete experiment configuration
  combined_summary.csv         # Final metrics for all runs
  dataset_name/                # e.g., adult, compas
    algorithm_name/            # e.g., fedavg, fairfed, fedcvg
      config_label/            # e.g., dir_alpha0.1_debias_true
        accuracy.csv           # Per-round accuracy values
        loss.csv               # Per-round loss values
        eod.csv                # Per-round EOD values
        spd.csv                # Per-round SPD values
        aod.csv                # Per-round AOD values
        precision.csv          # Per-round precision values
        config.json            # Run-specific configuration
        metrics_info.json      # Metadata about metrics
```

The `experiment_config.json` file at the root contains the complete experiment configuration, including all algorithm parameters, random seeds, and execution metadata, ensuring that every experiment can be exactly reproduced. The `combined_summary.csv` file aggregates final metrics (last round values) from all runs, facilitating comparative analysis across configurations. Each CSV file within a configuration directory contains time-series data with one row per training round, enabling detailed analysis of convergence behavior and fairness evolution during training.

Plots matching those presented in Chapter 5 can be generated using `create_experiment_plots.py`, which reads the CSV files and produces comparison charts for all metrics. Interrupted experiments can be resumed using the `--resume` flag, which automatically detects completed runs by checking for the presence of all expected CSV files and skips them. All sources of randomness—including data partitioning, model initialization, client selection, and training batch shuffling—are controlled by the seed parameter, guaranteeing identical results across runs.

Chapter 5

Experimental Results

This chapter presents the experimental results from evaluating the fairness-aware federated learning algorithms described in Chapter 3 under the experimental setup defined in Chapter 4. We conduct two main experiments to assess algorithm performance across different data partitioning strategies and heterogeneity conditions.

Section 5.1 presents Experiment 1, which evaluates algorithms under Dirichlet-based partitioning with varying heterogeneity levels ($\alpha \in \{0.1, 0.2, 0.5, 10.0, 5000.0\}$). This experiment analyzes how sensitive attribute imbalance across clients affects fairness-aware aggregation strategies. We examine baseline algorithm performance (Section 5.1.2), the impact of local debiasing (Section 5.1.3) and parity sampling (Section 5.1.4) as bias mitigation techniques, their interaction (Section 5.1.5), and combined performance metrics (Section 5.1.6).

Section 5.2 presents Experiment 2, which evaluates algorithms under coverage-based partitioning with explicit representation constraints. This experiment tests whether controlling client group representation directly improves fairness outcomes. We analyze three coverage levels for each dataset (Adult: 1999, 2570, 4497 samples per client; COMPAS: 474, 610, 1067 samples per client). The analysis follows the same structure as Experiment 1: baseline comparison (Section 5.2.2), local debiasing impact (Section 5.2.3), parity sampling impact (Section 5.2.4), technique interaction (Section 5.2.5), and combined metrics (Section 5.2.6).

Section 5.3 synthesizes findings from both experiments, providing algorithm family recommendations for different deployment scenarios based on data characteristics, heterogeneity levels, and operational constraints.

5.1 Experiment 1: Dirichlet Partitioning

The goal of the first experiment is to compare the reference algorithms with respect to different levels of client dataset heterogeneity, defined in terms of different Dirichlet partitioning, as described in Section 4.3.1. The considered Dirichlet partitionings span from highly non-IID distributions ($\alpha = 0.1$, where clients have very skewed sensitive attribute distributions) to effectively IID distributions ($\alpha = 5000$, where all clients have similar sensitive attribute distributions).

To this aim, we first provide information about client data distributions (Subsection 5.1.1) and we discuss how we manage coverage in FedCvg variants under Dirichlet partitioning. Then, we present and discuss the obtained results with the goal of analyzing baseline aggregation algorithms (Subsection 5.1.2), the impact of local debiasing (Subsection 5.1.3), the impact of parity sampling (Subsection 5.1.4), the interaction between local debiasing and parity sampling (Subsection 5.1.5), and the top approaches when considering the FAS combined metric (Subsection 5.1.6).

For all the algorithm versions described in Section 4.4, we consider one configuration for each combination of parameter values pointed out in Table 4.2 and all the metrics listed in Section 4.6 but FAS, which will be discussed in Subsection 5.1.6.

Tables A.1–A.5 in Appendix A present the comprehensive performance comparison on the Adult dataset while Tables B.1–B.5 in Appendix B present results for the COMPAS dataset, which has different characteristics (smaller size: 6,172 samples vs 48,842 for Adult; different sensitive attribute encoding: 81% male unprivileged (COMPAS) vs 33% female unprivileged (Adult)). In both cases, results are organized by heterogeneity level, with each table showing all the considered metrics for one α value. Following the FairFed paper methodology, we select the configuration corresponding to the best EOD value (closest to zero) across learning rates and report the mean across five random seeds. We observe that COMPAS consistently exhibits negative AccDiff values, indicating that the privileged group (female, 19% minority) attains substantially higher accuracy than the unprivileged group (male, 81% majority).

5.1.1 Client Data Distribution

Table 5.1 summarizes the resulting client distributions (average size per client; minimum, average, and maximum percentage of unprivileged samples; standard deviation) across the considered heterogeneity levels for both the considered datasets. Statistics are averaged over the 5 random seeds.

We observe the following:

Table 5.1: Client data distribution statistics under Dirichlet partitioning (averaged over 5 seeds)

Dataset	α	**Avg Size per Client**	**Min Unpriv (%)**	**Mean Unpriv (%)**	**Max Unpriv (%)**	**Std Unpriv (%)**
Adult Dataset (Global: 33% female unprivileged, 67% male privileged)						
	0.1	9,768	2%	33%	95%	38%
	0.2	9,768	5%	33%	82%	29%
	0.5	9,768	12%	33%	65%	20%
	10.0	9,768	22%	33%	45%	9%
	5000.0	9,768	31%	33%	35%	2%
COMPAS Dataset (Global: 81% male unprivileged, 19% female privileged)						
	0.1	1,234	8%	81%	99%	42%
	0.2	1,234	25%	81%	98%	31%
	0.5	1,234	48%	81%	95%	19%
	10.0	1,234	68%	81%	90%	8%
	5000.0	1,234	79%	81%	83%	2%

- *General trends*: For each α value, the client dataset size coincide in average and that, by construction, in average the average percentage of unprivileged samples correspond to the percentage of unprivileged samples in the dataset but the standard deviation decreases while decreasing the level of heterogeneity (i.e., for higher α values).

- *High heterogeneity ($\alpha = 0.1, 0.2$)*: Low α values create extreme representation imbalances. At $\alpha = 0.1$ on Adult, unprivileged representation (females) ranges from 2% to 95% across clients (std=38%), meaning some clients have almost no female samples while others have almost exclusively females. As we will see, this severe imbalance makes fairness particularly challenging.

- *Moderate heterogeneity ($\alpha = 0.5$)*: α values around 0.5 produce substantial but less extreme variation. On Adult, unprivileged representation (females) ranges from 12% to 65% (std=20%), ensuring all clients have meaningful representation of both groups while maintaining non-IID characteristics.

- *Near-IID ($\alpha = 10.0, 5000.0$)*: High α values leads to client distributions approaching the global one. At $\alpha = 5000$ on Adult, all clients have 31-35% female (unprivileged) representation (std=2%), essentially IID. This α setting serves as a control condition to isolate the effect of heterogeneity.

- *Dataset comparison*: COMPAS shows similar heterogeneity patterns but with higher baseline imbalance (81% male unprivileged vs 33% female unprivileged in Adult). This extreme global imbalance compounds the challenges created by client-level heterogeneity.

Unlike coverage-based partitioning where we explicitly designate good clients, Dirichlet partitioning creates a continuous spectrum. We can retrospectively identify clients with adequate representation (e.g., $\geq$30% of minority group), but this varies by seed and α. At $\alpha = 0.1$, typically 2-3 out of 5 clients have balanced representation, while at $\alpha = 5000$, all 5 clients are 'good'.

For the FedCvg execution a value for the coverage parameter *cov* should anyhow be selected. As explained in Section 4.4, when using Dirichlet partitioning, *cov* is set to the mean of the unprivileged sample count per client. This adaptive approach allows FedCvg to apply coverage-based weighting consistently across different heterogeneity levels. Table 5.2 shows the computed coverage values for each α level, averaged across the 5 random seeds.

Table 5.2: FedCvg coverage parameter (*cov*) under Dirichlet partitioning (mean unprivileged samples per client, averaged over 5 seeds)

Dataset	α	**Coverage Mean**	**Coverage Median (Std)**	**Good Clients (%)**
Adult Dataset (Total unprivileged: 16,117 females)				
	0.1	3,223	613 (992)	32%
	0.2	3,223	1,253 (1,251)	40%
	0.5	3,223	1,955 (1,225)	48%
	10.0	3,223	2,556 (268)	60%
	5000.0	3,223	2,584 (11)	60%
COMPAS Dataset (Total unprivileged: 4,999 males)				
	0.1	1,000	370 (228)	40%
	0.2	1,000	405 (124)	36%
	0.5	1,000	523 (99)	44%
	10.0	1,000	605 (59)	56%
	5000.0	1,000	616 (1)	60%

We observe the following:

- *Constant mean across α*: The coverage mean is constant across all α values because it represents the total number of unprivileged samples divided by the number of clients (5). For Adult, unprivileged are females (3,223 mean per client); for COMPAS, unprivileged are males (1,000 mean per client). This value serves as the baseline coverage threshold for FedCvg's weighting mechanism.

- *Median varies with heterogeneity*: The median unprivileged count per client varies dramatically with α, reflecting the actual distribution heterogeneity. At $\alpha = 0.1$ on Adult, the median is only 613 (mean 3,223), indicating that most clients have far fewer unprivileged samples than the mean due to extreme skew. At $\alpha = 5000$, median and mean converge (2,584 vs 3,223), confirming near-IID conditions.

- *Good clients increase with α*: The percentage of good clients (those with a number of unprivileged samples greater than the mean of unprivileged samples) increases with α. At $\alpha = 0.1$, only 32% of clients on Adult are good, reflecting extreme heterogeneity. At $\alpha = 5000$, 60% are good, approaching the theoretical 50% expected in IID conditions. COMPAS shows similar patterns, with good clients ranging from 36-40% at low α to 60% at high α. We observe however a reduction of good clients for $\alpha = 0.2$.

- *High variability at low α*: The standard deviation of the median across seeds is highest at low α values (992 for Adult at $\alpha = 0.1$), indicating that the exact distribution of unprivileged samples varies significantly across random seeds. This variability decreases as α increases, reaching near-zero at $\alpha = 5000$ (std=11).

- *Comparison with coverage-based partitioning*: For COMPAS, the coverage values used in coverage-based partitioning are comparable to the Dirichlet mean (1,000), but the key difference is that coverage-based partitioning explicitly ensures 50% of clients meet the threshold, while Dirichlet creates a continuous distribution where the number of good clients varies by seed and α (36-60% for COMPAS).

We expect that the client data distribution has an impact on the considered algorithms:

- FedAvg family: Since it weights clients by dataset size, it will give equal influence to clients regardless of their sensitive attribute distribution. For Adult, this means clients with 2% female (unprivileged) and 95% female receive equal per-sample weight, propagating representation bias to the global model.

- FairFed family: It adjusts weights based on fairness metrics computed on each client's data. Thus, clients with extreme imbalances may have undefined fairness metrics (e.g., no positive examples from one group), triggering the accuracy-gap fallback. For Adult at $\alpha = 0.1$, clients with only 2% females may lack sufficient samples for reliable fairness computation.

- FedCvg/FedCvg-Ratio family: They explicitly account for representation imbalances by boosting clients with better unprivileged representation. For Adult at $\alpha = 0.1$, clients with 95% female (unprivileged) will receive much higher weights than those with 2% female, directly addressing representation bias. For COMPAS, clients with more males (unprivileged) are favored.

- Parity Sampling: It can strategically select the 2-3 'good clients' when the differences among client data distributions are significant (e.g., at $\alpha = 0.1$), but it might become less effective for more homogeneous distributions (e.g., $\alpha = 5000$). For Adult, it selects clients with higher female representation; for COMPAS, clients with higher male representation.

- Local Debiasing: In this case, effectiveness depends on having sufficient samples for each combination of sensitive and target attributes. For Adult at $\alpha = 0.1$, clients with only 2% female representation may have too few female samples for effective reweighting.

5.1.2 Comparison of Aggregation Algorithms

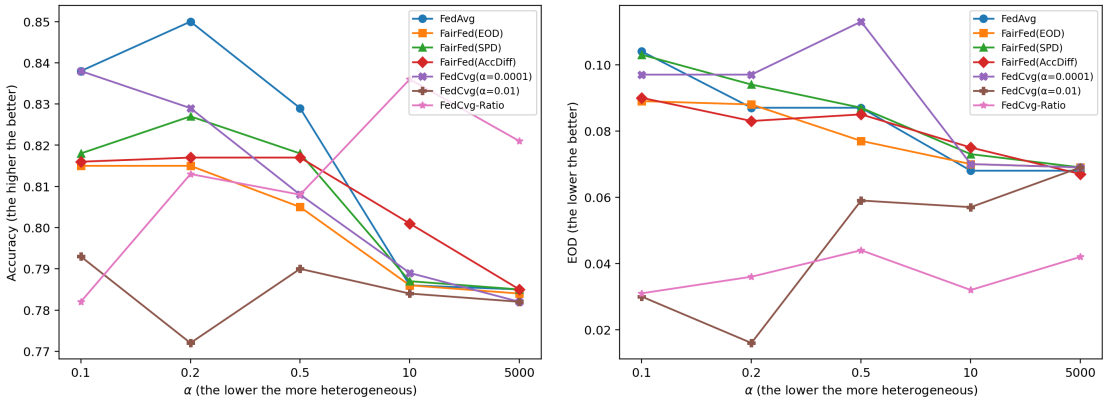
In this experiment, for each dataset, we compare the considered algorithm variants with respect to the heterogeneity level for all the metrics under analysis. For FedCvg-Ratio, the two configurations corresponding to $\lambda = 0.5$ and $\lambda = 0.9$ yield nearly indistinguishable results across all reported metrics; therefore, for clarity, only the $\lambda = 0.5$ variant is shown.

5.1.2.1 Adult Dataset

Figure 5.1 summarizes the baseline performance of the considered methods on the Adult dataset across different levels of heterogeneity α (the lower the more heterogeneous).

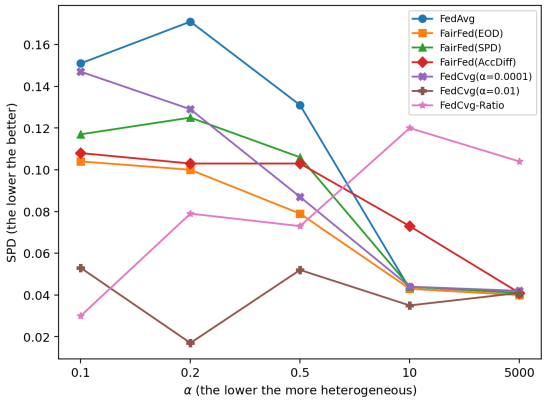
Trends comparison. From the plots, FedCvg is strongly influenced by the value of α_{cov}. With a small value ($\alpha = 0.0001$), its behavior is close to FedAvg and the FairFed variants in both fairness metrics and accuracy. When $\alpha = 0.01$, the impact of coverage becomes more pronounced, often improving fairness, especially under higher heterogeneity, at the cost of lower accuracy, thus highlighting a clearer trade-off. In this setting, Fed-Cvg shows trends very similar to FedCvg-Ratio, suggesting that a stronger coverage contribution makes its aggregation behavior closer to the ratio-based formulation. Overall, FedCvg-Ratio remains stable and achieves competitive fairness results while maintaining accuracy comparable to the other methods, representing a balanced compromise. The FairFed variants show similar overall trends across heterogeneity levels, with differences that reflect the specific fairness objective they target. However, no single variant consistently dominates the others on its corresponding metric, suggesting that the impact of the fairness constraint depends on the heterogeneity setting and the trade-off with accuracy.

Accuracy. In terms of predictive performance (Fig. 5.1(a)), FedAvg achieves the highest accuracy under high and moderate heterogeneity (up to $\alpha = 0.5$), reaching 0.850 at $\alpha = 0.2$, and 0.838 at $\alpha = 0.1$. In the same regimes, Fed-Cvg(0.01) and FedCvg-Ratio achieve the worst results, especially for $\alpha = 0.1$. As α increases and data become more homogeneous, FedCvg-Ratio becomes competitive, while FedCvg(0.01) accuracy remains low. FairFed variants exhibit slightly lower but relatively stable accuracy across all regimes. These results confirm that, at the baseline level, FedAvg is primarily optimized for predictive performance, particularly under high or moderate heterogeneity, whereas representation-aware methods can match or closely approach its accuracy in near-IID con-

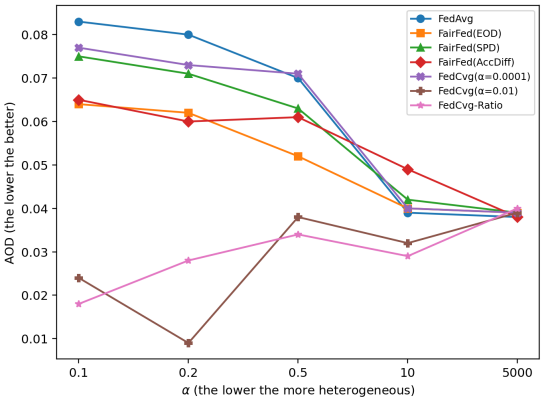


(a) Accuracy

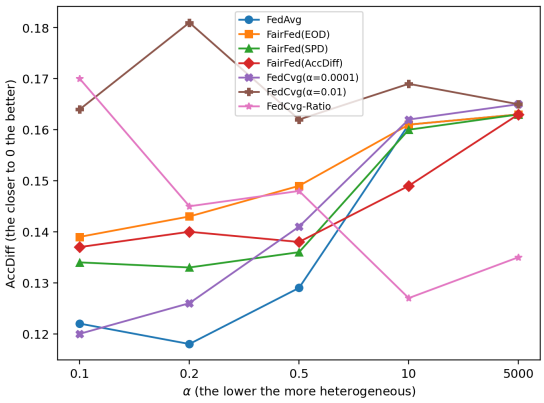
(b) EOD



(c) SPD



(d) AOD



(e) AccDiff

Figure 5.1: Baseline performance on the Adult dataset across heterogeneity levels α (the lower the more heterogeneous).

ditions.

Fairness metrics. Fairness trends are particularly informative. In Fig. 5.1(b), FedCvg(0.01) and FedCvg-Ratio achieve markedly lower EOD values under strong non-IID conditions compared to the other approaches. For instance, at $\alpha = 0.1$, FedAvg reports EOD= 0.104, while FedCvg-Ratio reduces it to 0.031, corresponding to an improvement of over 70%. Similar gaps are observed at $\alpha = 0.2$. The FairFed variants improve over FedAvg in several heterogeneous settings, but their gains are generally more moderate and less consistent than those of representation-based methods. Analogous patterns emerge for SPD and AOD, where representation-based aggregation substantially mitigates disparities in highly heterogeneous regimes. As α increases and client distributions approach IID, these differences progressively shrink, indicating that the corrective effect of representation-aware aggregation is most pronounced under strong heterogeneity. While representation-based variants generally achieve the most favorable results in EOD and AOD, their advantage becomes less evident in near-IID settings; the FairFed family, in contrast, exhibits more stable but comparatively smaller fairness improvements across heterogeneity levels, with values between those of FedAvg and representation-based approaches.

Accuracy difference. The AccDiff metric (Fig. 5.1(e)) shows a different behavior: by increasing α, thus moving towards IID distributions, the difference increases (thus, disparity increases) for all techniques except FedCvg(0.01) and FedCvg-Ratio, with the second one leading to the best results for homogeneous settings. This suggests that parity in predictive accuracy across groups is achieved through mechanisms partially distinct from those governing fairness metrics.

Conclusions. The results show three distinct behaviors. FedAvg achieves the highest accuracy, particularly under high and moderate heterogeneous settings, but as expected exhibits larger fairness disparities on the Adult dataset, which is inherently imbalanced across target classes. The FairFed variants provide moderate and more stable fairness improvements while preserving competitive accuracy, yet without consistently dominating across metrics. In contrast, representation-based methods, especially FedCvg-Ratio, deliver the strongest fairness gains under non-IID conditions and maintain clear advantages in AccDiff, highlighting the effectiveness of representation-aware aggregation in mitigating disparities induced by skewed data distributions. As α increases and data become more homogeneous, differences between methods progressively shrink, and disparities naturally decrease. In this case, representation-based aggregation seems to primarily act as a corrective mechanism in highly heterogeneous regimes.

5.1.2.2 COMPAS Dataset

Figure 5.2 summarizes the baseline performance of the considered methods on the COMPAS dataset across different levels of heterogeneity α.

Trend comparison. From the COMPAS plots, FedCvg remains influenced by α_{cov}, but the difference between $\alpha = 0.0001$ and $\alpha = 0.01$ is much less pronounced than in Adult. The two variants show very similar trends across both accuracy and fairness metrics. Moreover, their behavior is still close to that of FedCvg-Ratio, suggesting that on COMPAS the strength of the coverage term does not substantially change the aggregation dynamics. FedCvg-Ratio for this dataset guarantees the best values for accuracy and fairness metrics, almost consistently with respect to heterogeneity levels, but leads to the worst results for the accuracy difference, especially in highly heterogeneous environments. As expected, FedAvg remains competitive only when considering accuracy. The FairFed family provides moderate and relatively stable fairness improvements over FedAvg, typically with limited accuracy degradation. However, as in Adult, no single FairFed variant clearly dominates on its target metric. Overall, on COMPAS the differences between FedAvg, FairFed, and representation-based approaches are less sharply separated.

Accuracy. Accuracy does not exhibit a uniform trend with respect to α. For all α values, FedCvg-Ratio guarantees the highest accuracy. This is especially evident for strong heterogeneity ($\alpha = 0.1$ and 0.2) scenarios. For $\alpha = 0.1$, FedCvg-Ratio and FairFed(EOD) have similar trends but for $\alpha = 0.2$ FairFed(EOD) significantly downgrades. At $\alpha = 0.5$, all the methods become comparable, with FedAvg slightly outperforming representation-based approaches. When data becomes more homogeneous, FedCvg-Ratio and FedCvg(0.0001) have similar trends and improve over the other approaches. This trend becomes more evident in the near-IID regime ($\alpha = 5000$). Overall, predictive performance does not simply converge across methods as heterogeneity decreases; instead, representation-aware aggregation becomes increasingly competitive, and eventually superior, in more homogeneous regimes.

Fairness metrics. Fairness behavior is strongly regime-dependent. Under high heterogeneity ($\alpha = 0.1$), representation-based aggregation substantially reduces disparity: for instance, FedCvg-Ratio halves EOD compared to FedAvg (0.029 vs 0.059), and similar improvements are observed for SPD and AOD. This advantage persists at $\alpha = 0.2$ and remains visible at $\alpha = 0.5$. However, a clear inversion appears at $\alpha = 10$, where representation-aware methods exhibit significantly larger disparity than FedAvg and the FairFed family across all fairness metrics. In the near-IID setting ($\alpha = 5000$), fairness gaps remain higher for representation-based strategies. Hence, unlike what might be expected, fairness does not monotonically improve as α increases; instead, representation-aware aggregation is particularly effective in highly heterogeneous settings, while its relative benefit diminishes—and may reverse—in more homogeneous regimes.

Accuracy difference. The behavior of AccDiff further highlights the non-monotonic nature of the trends. For $\alpha = 0.1$, 0.2, and 0.5, representation-aware methods generally yield larger accuracy gaps between groups than the other approaches, with worst results achieved by FedCvg-Ratio. At $\alpha = 10$, however, representation-based approaches like FedCvg-

RATIO achieve a smaller disparity. This advantage does not fully persist at $\alpha = 5000$, where differences become comparable again. Therefore, AccDiff follows a pattern partially distinct from EOD, SPD, and AOD, suggesting that mechanisms affecting classification accuracy across groups are not perfectly aligned with those governing error-rate disparities.

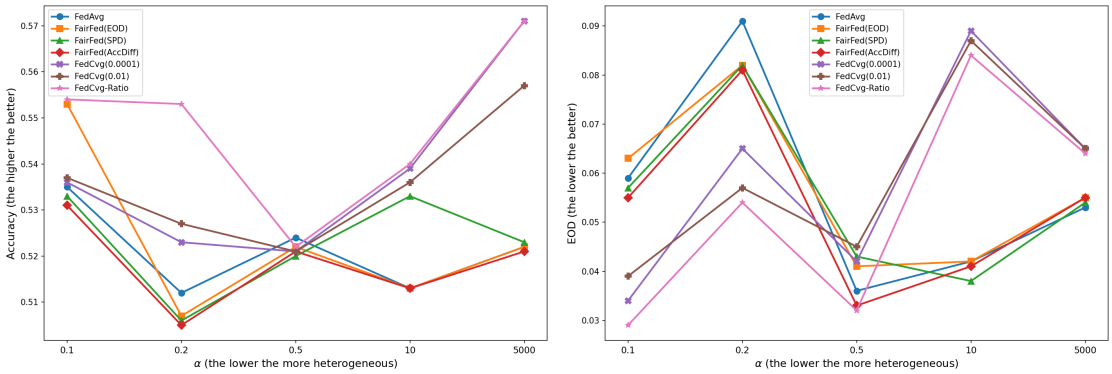
Conclusions. The analysis shows that on COMPAS the impact of the aggregation strategy is highly heterogeneity-dependent. The FEDCVG variants behave very similarly to each other and close to FEDCVG-RATIO, indicating that the strength of the coverage term plays a limited role on this dataset. Representation-aware methods provide clear fairness gains under strong non-IID conditions and remain competitive in accuracy, especially FEDCVG-RATIO, but may incur larger fairness or accuracy gaps in intermediate-to-high α regimes. FEDAVG remains mainly accuracy-oriented, while the FAIRFED family offers moderate and more stable fairness improvements without clearly dominating across settings. These regime-dependent results are likely influenced by the intrinsic properties of COMPAS and the choice of the majority group.

5.1.3 Impact of Local Debiasing

Figures 5.3 and 5.4 report the impact of Local Debiasing (LD) across different heterogeneity levels on the Adult dataset. Each bar in the histograms represents the variation $\Delta = (\mathcal{A} + LD) - \mathcal{A}$ for a specific metric, where $\mathcal{A}$ is one of the considered algorithms. Negative values indicate an improvement for fairness metrics (SPD, EOD, AOD, AccDiff), while positive values indicate an improvement in Accuracy. For what concerns FAIRFED, for SPD, EOD, and AccDiff we report only the variant trained on that metric, to simplify the analysis.

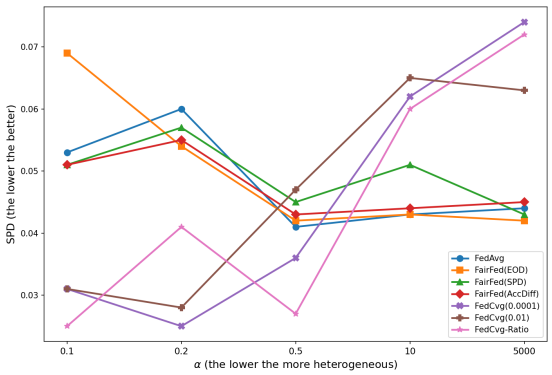
5.1.3.1 Adult Dataset

Trend comparison. First of all, we observe that for all the metrics, variations are very low, under 0.05. However, the emerging pattern is that LD systematically improves fairness-related metrics. Indeed, EOD, SPD, and AOD exhibit predominantly negative Δ values for any heterogeneity level, indicating lower disparity after introducing LD. The impact on AccDiff and accuracy are less evident but also in this case better results are achieved for low α values. In those cases, as expected, the magnitude of the improvement is typically modest compared to the gains observed in fairness metrics. The highest improvements are achieved by FEDCVG(0.0001), especially in high and moderate heterogeneous environments, while FEDCVG(0.01) and FEDCVG-RATIO achievements are less evident, especially at high heterogeneity levels, probably because in those situations local debiasing limits the effect of coverage and ratio-based approaches. The impact of local

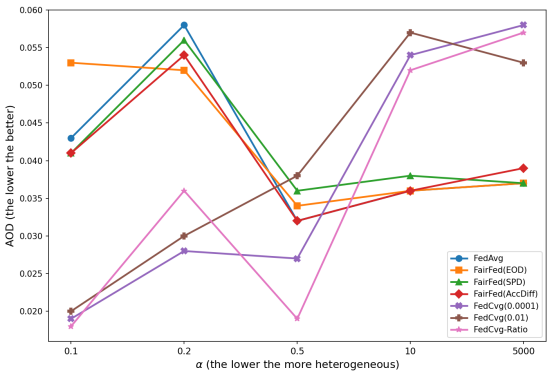


(a) Accuracy

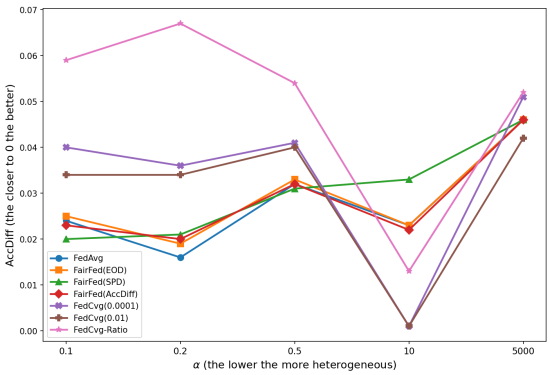
(b) EOD



(c) SPD



(d) AOD



(e) AccDiff

Figure 5.2: Baseline performance on the COMPAS dataset across heterogeneity levels α (the lower the more heterogeneous).

debiasing on FedAvg and the FairFed family is quite similar, especially at moderate and low heterogenous levels.

High heterogeneity. Under strong non-IID conditions, LD produces the largest fairness gains. In particular, for $\alpha = 0.1$, reductions in EOD and AOD are particularly pronounced across most techniques. The improvement in fairness is often accompanied by a slight accuracy improvement. For $\alpha = 0.2$, the behavior remains similar but slightly attenuated. Fairness gains persist while the impact on accuracy is generally smaller or null.

Moderate heterogeneity. At intermediate heterogeneity, LD still improves fairness metrics, though the magnitude of Δ decreases compared to the highly non-IID case. Interestingly, some methods (e.g., FedCvg(0.0001)) significantly improve EOD and AOD and accuracy at the same time, suggesting that LD can act as a stabilizing regularizer when data heterogeneity is moderate.

Low heterogeneity. When the data distribution approaches IID, the effect of LD becomes less pronounced for all techniques. Fairness improvements are smaller in magnitude (with the higher improvement achieved by FedCvg methods) and the variations across methods tend to shrink. This indicates that LD is most beneficial in heterogeneous settings, where disparities across clients are stronger.

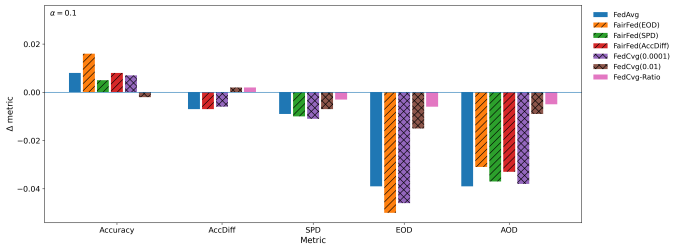
Conclusions. The results suggest that LD is particularly effective as a fairness-enhancing mechanism (especially on EOD, and to a lesser extent AOD) in highly non-IID federated settings, with limited and controlled impact on accuracy. As α increases and the setting approaches IID, the impact of LD becomes progressively negligible. The effects are less evident on techniques relying on coverage or ratio-based aggregation (FedCvg(0.01) and FedCvg-Ratio), demonstrating that representation-based server-side approaches represent in this case a valuable alternative to client-side debiasing techniques.

We observe that these results are expected considering the characteristics of the Adult dataset and the choice of the unpriviled group. Indeed, in Adult, women are significantly underrepresented in the favorable outcome (income $\geq 50K$), producing a clear imbalance in the joint distribution of gender and label. Since reweighing directly adjusts this distribution to reduce dependence between the sensitive attribute and the outcome, it effectively mitigates demographic disparities.

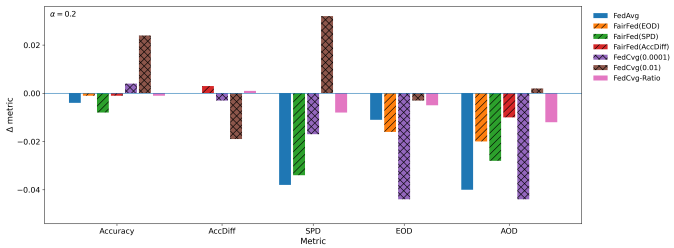
5.1.3.2 COMPAS Dataset

Trend comparison. For the COMPAS dataset, variations are mostly around 0, with only few exceptions related to different techniques for different heterogeneous levels. No significant trends are identified.

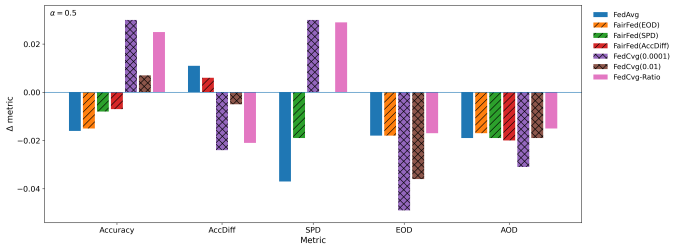
High heterogeneity. Under strong heterogeneity ($\alpha = 0.1$ and $\alpha = 0.2$), LD produces



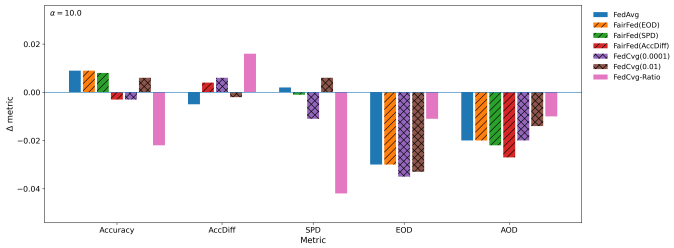
(a) $\alpha = 0.1$



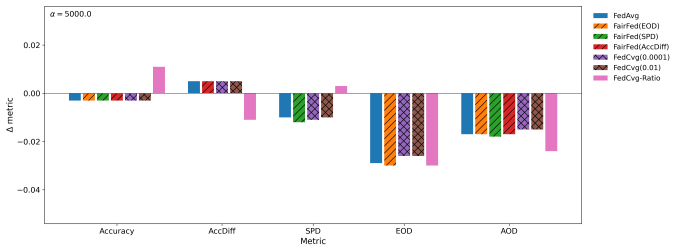
(b) $\alpha = 0.2$



(c) $\alpha = 0.5$



(d) $\alpha = 10$



(e) $\alpha = 5000$

Figure 5.3: Impact of LD across heterogeneity levels on the Adult dataset.

mixed and metric-dependent effects but overall only few improvements are observed. More precisely, accuracy generally remains stable or slightly decreases at $\alpha = 0.1$, with a more visible degradation for FairFed(EOD). Regarding fairness and accuracy difference, LD improves the metrics only in few specific cases; for example, negative values for SPD or AOD are observed in certain representation-based variants. We observe that, for $\alpha = 2$, the LD effects are more evident on FedCvg-Ratio, with main benefits on accuracy and fairness metrics.

Moderate heterogeneity. At intermediate heterogeneity ($\alpha = 0.5$), the impact of LD is very limited. Accuracy variations are close to zero across methods, and fairness metrics exhibit only marginal changes, with small positive and negative Δ values that do not reveal a clear systematic trend. This suggests that in moderately heterogeneous regimes the baseline optimization is already relatively stable and LD does not substantially reshape the fairness–accuracy balance.

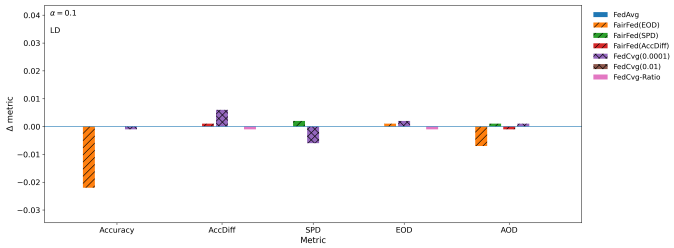
Low heterogeneity. In low-heterogeneity and IID regimes ($\alpha = 10$ and $\alpha = 5000$), LD effects become more evident but remain method-dependent. Accuracy changes are mostly small and occasionally positive. For some representation-based configurations at $\alpha = 10$, AccDiff decreases significantly. However, this often coincides with positive Δ values in SPD, EOD, or AOD. At $\alpha = 5000$, most Δ values are close to zero, showing that LD has minimal influence when client distributions are fully homogeneous.

Conclusions. LD does not uniformly improve fairness across heterogeneity levels. The effects are irregular and metric-dependent. This results are in line with the characteristics of the COMPAS dataset when gender is treated as the sensitive attribute. Indeed, since local debiasing methods mainly enforce demographic parity, while COMPAS fairness tensions concern error-rate trade-offs, adjusting instance weights produces only marginal changes in the model.

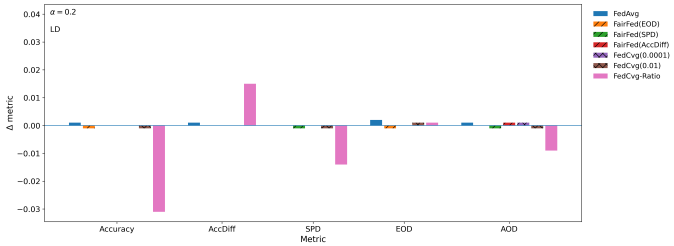
5.1.4 Impact of Parity Sampling

5.1.4.1 Adult Dataset

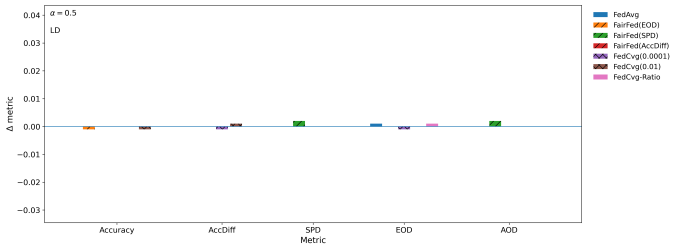
Trend comparison. The introduction of *parity sampling* (PS) has in general a limited impact (variations up to 0.03), with either a positive or negative impact. The highest benefits on fairness metrics are obtained with FedAvg, where the impact of PS is mostly positive and higher for higher heterogeneity level, at the price of a lower accuracy. For FedCvg variants and FedCvg-Ratio, the impact of PS often becomes positive for near-IID configurations. The behavior of FairFed variants is more heterogeneous and depends on the considered metric and the heterogeneity level. For all approaches, the impact is mostly null for IID configurations.



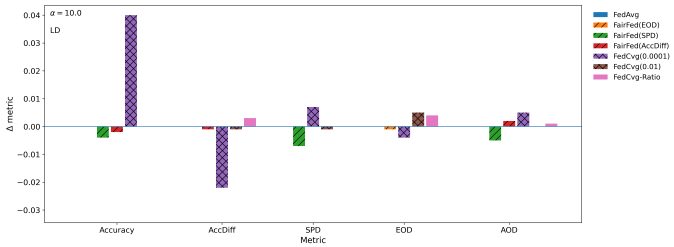
(a) $\alpha = 0.1$



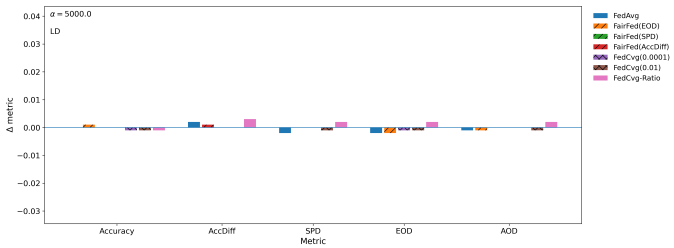
(b) $\alpha = 0.2$



(c) $\alpha = 0.5$



(d) $\alpha = 10$



(e) $\alpha = 5000$

Figure 5.4: Impact of LD across heterogeneity levels on the COMPAS dataset.

High heterogeneity. In the high-heterogeneity regime ($\alpha = 0.1$ and $\alpha = 0.2$), the impact of PS is often negative. When $p = 0.7$, the effects are even more pronounced: some techniques of the FairFed family show accuracy improvements ($\Delta > 0$), but often accompanied by deteriorations in fairness metrics (positive Δ). The behavior varies significantly across methods and FedCvg(0.01) and FedCvg-Ratio exhibit larger oscillations in fairness metrics.

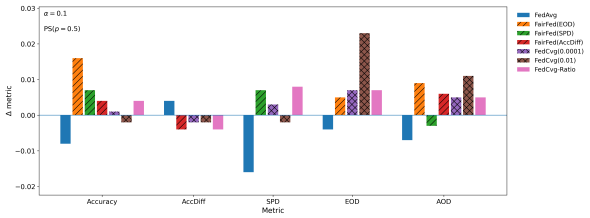
Moderate heterogeneity. With intermediate heterogeneity ($\alpha = 0.5$), the effect of PS appears more balanced. For $p = 0.5$, some methods slightly improve fairness metrics, while accuracy changes remain limited (often slightly negative for FedAvg). With $p = 0.7$, differences become more noticeable but less extreme than in the high-heterogeneity case. In this regime, PS can yield selective improvements without systematically degrading the other metrics. Overall, medium heterogeneity seems to provide the most stable trade-off between accuracy and fairness under PS.

Low heterogeneity. Under low heterogeneity ($\alpha = 10$ and $\alpha = 5000$), the differences between configurations with and without PS are minimal. For $\alpha = 5000$, almost all metrics exhibit Δ values close to zero, for both $p = 0.5$ and $p = 0.7$, indicating that PS is essentially neutral. Similarly, for $\alpha = 10$, variations remain small, although minor fluctuations can be observed. In this regime, the trade-off between accuracy and fairness is weak, and the methods behave similarly. Therefore, the benefit of PS is marginal when the client data distribution is already relatively homogeneous.

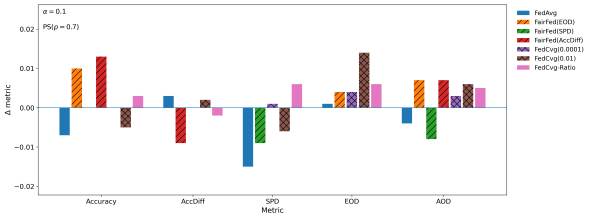
Conclusions. The benefit of using PS is less evident than using LD, often leading to worst values compared to the baselines. Better trade-offs are observed for medium heterogeneity. The most significant positive impact on fairness metrics is achieved by techniques which do not apply any specific bias-oriented aggregation method, like FedAvg, particularly under high heterogeneity. Fairness improvements come at the price of accuracy downgrade. Increasing p from 0.5 to 0.7 generally amplifies the PS effect. The most affected metric is SPD which. This is expected since PS try to amplify statistical parity. As α increases, the effect of PS progressively vanishes. This is expected since for more homogeneous samples parity sampling is not very effective.

5.1.4.2 COMPAS Dataset

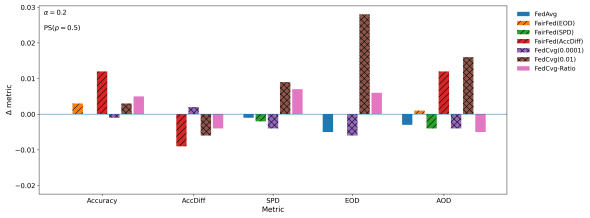
Trend comparison. With COMPAS, the impact of PS is even more limited than with Adult. We notice a generalized downgrade of accuracy for FedAvg, with some exceptions for $p = 0.7$ and high heterogeneity. At the same time, the impact on EOD and AOD for all techniques and all heterogeneity levels is highly variable and clear trends cannot be identified. At the same time, we notice however a generalized positive impact (thus a reduction) on the accuracy difference and SPD for almost all techniques and heterogeneity levels.



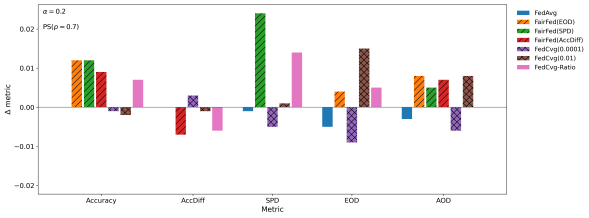
(a) $\alpha = 0.1$, $p = 0.5$



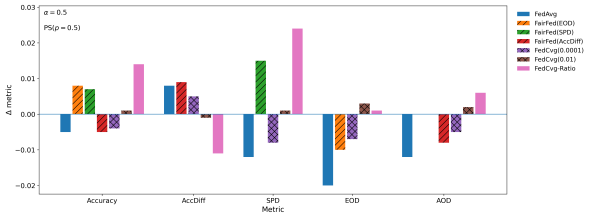
(b) $\alpha = 0.1$, $p = 0.7$



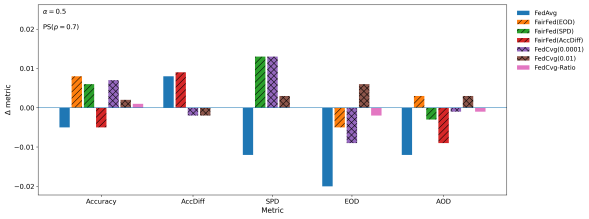
(c) $\alpha = 0.2$, $p = 0.5$



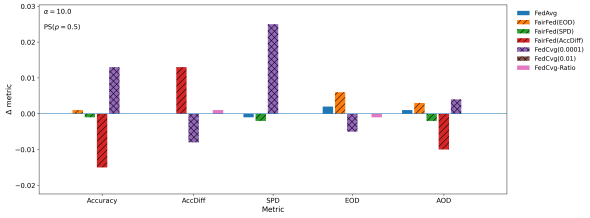
(d) $\alpha = 0.2$, $p = 0.7$



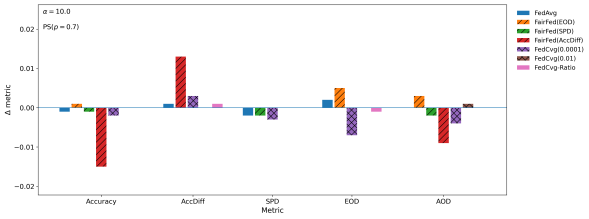
(e) $\alpha = 0.5$, $p = 0.5$



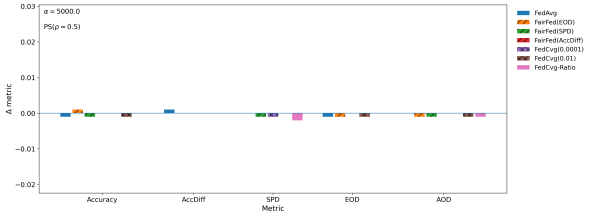
(f) $\alpha = 0.5$, $p = 0.7$



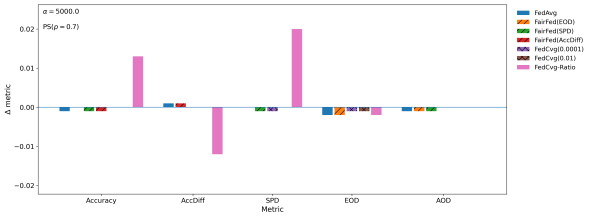
(g) $\alpha = 10$, $p = 0.5$



(h) $\alpha = 10$, $p = 0.7$



(i) $\alpha = 5000$, $p = 0.5$



(j) $\alpha = 5000$, $p = 0.7$

Figure 5.5: Impact of PS across heterogeneity levels on the Adult dataset.

High heterogeneity. In high heterogeneity, PS acts as a moderate corrective mechanism: it slightly adjusts fairness gaps, with limited impact on accuracy, and interacts more visibly with methods that do not already encode strong fairness constraints. For FedAvg, PS slightly improves fairness metrics at $\alpha = 0.2$ but at the same time accuracy downgrades. The FairFed family exhibits more variable behavior, probably because. As already observed, this indicates that when fairness constraints are already embedded in the optimization, PS acts as a secondary adjustment rather than a primary driver of fairness gains. For FedCvg methods, PS may slightly improve some fairness gaps but can also introduce small trade-offs in accuracy. The FedCvg-Ratio variants tend to show very limited sensitivity to PS in this regime, suggesting that their ratio-based reweighting already dominates the fairness–accuracy trade-off.

Moderate heterogeneity. At intermediate heterogeneity ($\alpha = 0.5$), FedCvg variants and FedCvg-Ratio guarantees slightly better fairness and accuracy difference values The impact on the other techniques is marginal.

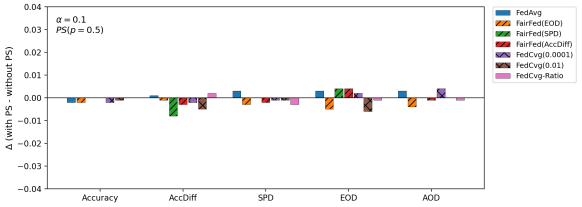
Low heterogeneity. For high α values, we notice some improvements on accuracy and accuracy difference for $\alpha = 10$ No other clear trend emerges.

Conclusions. On COMPAS the effect of PS is even more irregular and metric-dependent across all heterogeneity regimes, often below 0.01. This behavior might be due to the structural properties of COMPAS, which is smaller and characterized by extreme sensitive attribute imbalance but balanced classes, and the choice of the unprivileged group. In such a setting, altering the sampling probability directly affects group representation, leading to heterogeneous and less predictable outcomes. Overall, while a larger p acts as a stronger but relatively predictable fairness intervention on Adult, on COMPAS it behaves more as a perturbation factor, increasing variability without consistently improving disparity metrics.

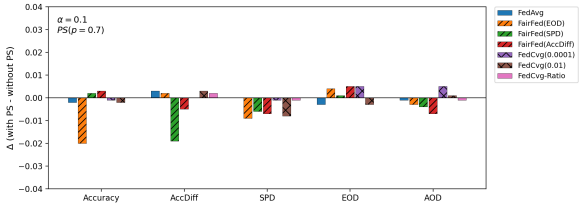
5.1.5 Interaction of Local Debiasing with Parity Sampling

After having discussed the impact of LD and PS separately, we now discuss their joint effect. To this aim, Table 5.3 presents a comprehensive analysis of EOD values at $\alpha = 0.1, 0.5, 5000$ for both Adult and COMPAS datasets, including all algorithm variants. We considered EOD as reference metric since the results reported in the tables in the appendix corresponds to the configurations leading to the best EOD values among all the considered learning rates.

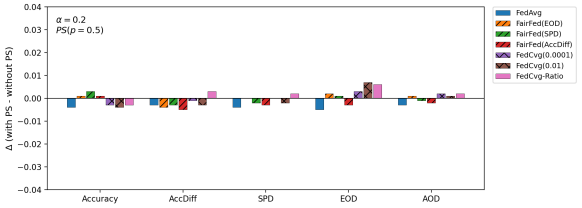
Adult dataset. On Adult, the LD+PS combination is more effective for algorithms where LD already induces a substantial fairness improvement. At $\alpha = 0.1$, FedAvg reduces EOD from 0.104 (Baseline) to 0.065 with LD, and further to 0.062 with LD+PS, indicating a small but consistent complementary effect. A similar pattern holds for



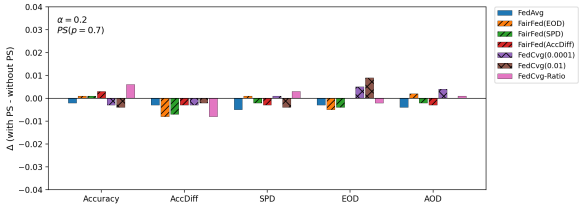
(a) $\alpha = 0.1$, $p = 0.5$



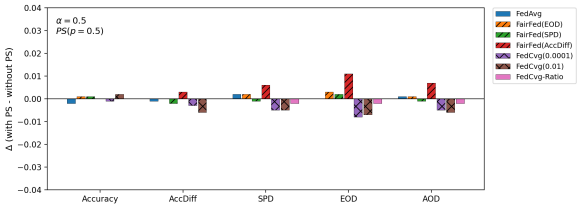
(b) $\alpha = 0.1$, $p = 0.7$



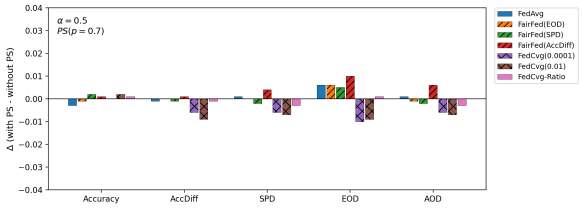
(c) $\alpha = 0.2$, $p = 0.5$



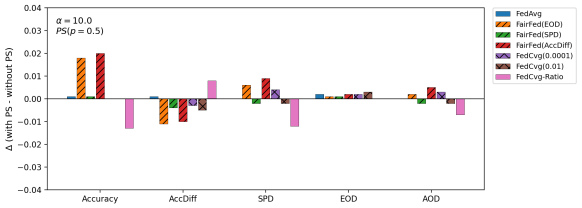
(d) $\alpha = 0.2$, $p = 0.7$



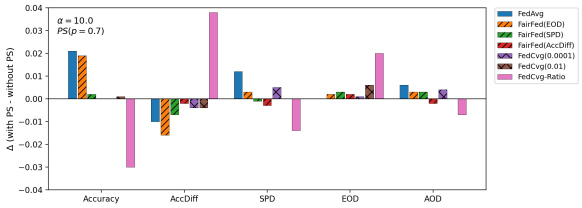
(e) $\alpha = 0.5$, $p = 0.5$



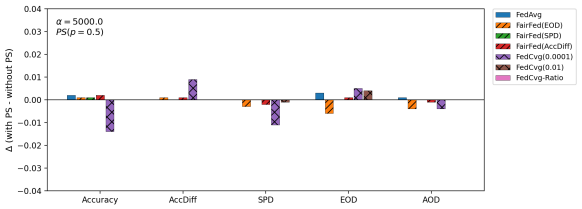
(f) $\alpha = 0.5$, $p = 0.7$



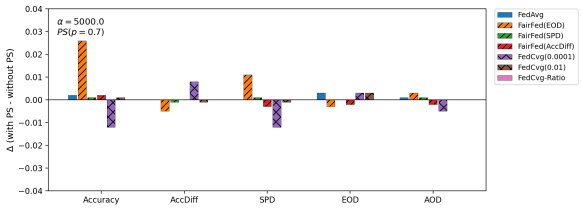
(g) $\alpha = 10$, $p = 0.5$



(h) $\alpha = 10$, $p = 0.7$



(i) $\alpha = 5000$, $p = 0.5$



(j) $\alpha = 5000$, $p = 0.7$

Figure 5.6: Impact of PS across heterogeneity levels on the COMPAS dataset.

Table 5.3: Effect of Local Debiasing and Parity Sampling combinations at $\alpha = 0.1, 0.5, 5000$ (EOD values)

$\alpha = 0.1$								
	Adult Dataset				**COMPAS Dataset**			
Algorithm	**Baseline**	**PS**	**LD**	**LD+PS**	**Baseline**	**PS**	**LD**	**LD+PS**
FedAvg	0.104	0.100	0.065	0.062	0.059	0.056	0.059	0.060
FedCvg(0.0001)	0.097	0.101	0.051	0.047	0.034	0.036	0.036	0.035
FedCvg(0.01)	0.030	0.044	0.015	0.018	0.039	0.033	0.039	0.035
FedCvg-Ratio(0.5)	0.031	0.037	0.025	0.019	0.029	0.028	0.028	0.029
FedCvg-Ratio(0.9)	0.031	0.035	0.025	0.026	0.029	0.027	0.028	0.027
FairFed(EOD)	0.089	0.093	0.039	0.031	0.063	0.058	0.064	0.061
FairFed(AccDiff)	0.090	0.094	0.042	0.036	0.055	0.059	0.055	0.058
FairFed(SPD)	0.103	0.094	0.052	0.050	0.057	0.058	0.056	0.058
$\alpha = 0.5$								
	Adult Dataset				**COMPAS Dataset**			
Algorithm	**Baseline**	**PS**	**LD**	**LD+PS**	**Baseline**	**PS**	**LD**	**LD+PS**
FedAvg	0.087	0.067	0.069	0.048	0.036	0.042	0.037	0.037
FedCvg(0.0001)	0.113	0.104	0.064	0.056	0.042	0.034	0.041	0.033
FedCvg(0.01)	0.059	0.065	0.023	0.030	0.045	0.038	0.045	0.035
FedCvg-Ratio(0.5)	0.044	0.042	0.027	0.015	0.032	0.030	0.033	0.030
FedCvg-Ratio(0.9)	0.044	0.041	0.027	0.016	0.032	0.031	0.033	0.030
FairFed(EOD)	0.077	0.072	0.059	0.043	0.041	0.047	0.041	0.047
FairFed(AccDiff)	0.085	0.066	0.053	0.040	0.033	0.043	0.033	0.045
FairFed(SPD)	0.087	0.072	0.057	0.045	0.043	0.048	0.039	0.047
$\alpha = 5000$								
	Adult Dataset				**COMPAS Dataset**			
Algorithm	**Baseline**	**PS**	**LD**	**LD+PS**	**Baseline**	**PS**	**LD**	**LD+PS**
FedAvg	0.068	0.066	0.039	0.038	0.053	0.056	0.051	0.055
FedCvg(0.0001)	0.069	0.068	0.043	0.043	0.065	0.070	0.064	0.061
FedCvg(0.01)	0.069	0.068	0.043	0.044	0.065	0.069	0.064	0.062
FedCvg-Ratio(0.5)	0.042	0.031	0.012	0.016	0.064	0.063	0.066	0.062
FedCvg-Ratio(0.9)	0.042	0.032	0.012	0.014	0.064	0.065	0.066	0.063
FairFed(EOD)	0.069	0.068	0.039	0.039	0.055	0.052	0.053	0.052
FairFed(AccDiff)	0.067	0.068	0.038	0.038	0.055	0.053	0.054	0.055
FairFed(SPD)	0.069	0.067	0.037	0.038	0.054	0.054	0.054	0.051

FairFed(EOD), which moves from 0.089 (Baseline) to 0.039 with LD and further to 0.031 under LD+PS. Here, PS amplifies the corrective effect of LD, suggesting that parity sampling refines the local representation adjustments introduced by LD. The strongest synergy appears in fairness-oriented methods such as FairFed(EOD) and FairFed(AccDiff), where LD already aligns model updates toward parity and PS further reduces residual imbalance. Moderate synergy is also visible for FedCvg(0.0001) and FedCvg(0.01), especially at $\alpha = 0.1$ and $\alpha = 0.5$, where LD+PS consistently yields slightly lower EOD than LD alone (e.g., $0.064 \rightarrow 0.056$ at $\alpha = 0.5$ for FedCvg(0.0001)). In contrast, the synergy is weaker for FedCvg-Ratio variants. Although LD significantly improves fairness (e.g., $0.031 \rightarrow 0.025$ at $\alpha = 0.1$ for $\lambda = 0.5$), the addition of PS produces only marginal adjustments (0.019–0.026 range). This suggests that ratio-based reweighting already internalizes most of the fairness correction, leaving limited room for PS to provide additional gains.

At intermediate heterogeneity ($\alpha = 0.5$), the LD+PS combination generally tracks LD alone across families, with the largest relative gains still observed for FedAvg and FairFed variants. Under near-IID conditions ($\alpha = 5000$), synergy becomes negligible across all techniques. In this regime, fairness disparities are already stable, and both LD and PS operate near their performance floor, leading to near-identical EOD values between LD and LD+PS.

COMPAS dataset. On COMPAS, the interaction between LD and PS is markedly less effective and more technique-dependent. At $\alpha = 0.1$, FedAvg shows virtually no benefit from LD (EOD remains 0.059) and LD+PS slightly worsens it (0.060), indicating the absence of complementarity. For FairFed(EOD), LD increases EOD ($0.063 \rightarrow 0.064$), and LD+PS partially recovers (0.061), but does not outperform the baseline configuration, highlighting a corrective rather than synergistic interaction.

Among FedCvg methods, synergy is minimal and inconsistent. For example, at $\alpha = 0.2$, FedCvg(0.01) improves under PS ($0.039 \rightarrow 0.033$), but LD does not reinforce this trend, and LD+PS remains close to LD alone. Similarly, FedCvg-Ratio variants display negligible or method-specific shifts under LD+PS, often mirroring LD with differences in the third decimal place.

At $\alpha = 0.5$, some localized gains appear (e.g., FedCvg-Ratio($\lambda = 0.5$) moves from 0.032 to 0.030 under LD+PS), but these changes are small and do not indicate systematic synergy across the family. Under near-IID conditions ($\alpha = 5000$), the interaction is largely neutral: LD+PS either matches LD or produces marginal fluctuations, without consistent amplification of fairness improvements.

Overall, the least synergy is observed for FedAvg and FedCvg variants on COMPAS, while limited and occasional complementarity is visible only in certain fairness-oriented configurations.

Table 5.4: Top 5 methods by Combined FAS (Adult dataset, Dirichlet partitioning)

α	Method	Acc	EOD	SPD	AOD	AccDiff	FAS
High Heterogeneity ($\alpha = 0.1$)							
	FedCvg(0.0001)+LD+PS(0.5)	0.845	0.047	0.140	0.039	0.114	**0.773**
	FedCvg(0.0001)+LD	0.845	0.051	0.136	0.039	0.114	0.773
	FedCvg(0.0001)+LD+PS(0.7)	0.844	0.051	0.136	0.037	0.115	0.773
	FairFed(AccDiff)+LD+PS(0.5)	0.834	0.036	0.113	0.026	0.125	0.771
	FedAvg+LD+PS(0.7)	0.846	0.062	0.140	0.042	0.115	0.770
Moderate Heterogeneity ($\alpha = 0.5$)							
	FedCvg-Ratio(0.5)+LD+PS(0.5)	0.833	0.024	0.099	0.017	0.126	**0.778**
	FedCvg-Ratio(0.9)+LD+PS(0.7)	0.833	0.024	0.102	0.019	0.126	0.777
	FedCvg-Ratio(0.5)+LD+PS(0.7)	0.833	0.026	0.100	0.018	0.127	0.777
	FedCvg-Ratio(0.5)+LD	0.833	0.027	0.102	0.019	0.127	0.776
	FedCvg-Ratio(0.9)+LD	0.833	0.027	0.102	0.019	0.127	0.776
Low Heterogeneity ($\alpha = 5000$)							
	FedCvg-Ratio(0.5)+LD+PS(0.5)	0.842	0.016	0.124	0.019	0.115	**0.784**
	FedCvg-Ratio(0.9)+LD+PS(0.5)	0.832	0.013	0.105	0.014	0.125	0.778
	FedCvg-Ratio(0.5)+LD	0.832	0.012	0.107	0.016	0.124	0.778
	FedCvg-Ratio(0.9)+LD	0.832	0.012	0.107	0.016	0.124	0.778
	FedCvg-Ratio(0.9)+LD+PS(0.7)	0.832	0.016	0.105	0.014	0.125	0.778

Conclusions. Across datasets, synergy is strongest for techniques where LD already produces substantial fairness gains and where the aggregation rule explicitly incorporates fairness objectives. This is evident in Adult for FairFed(EOD) and, to a lesser extent, FedAvg. Synergy is moderate for FedCvg methods and weakest for FedCvg-Ratio, whose internal reweighting mechanism appears to subsume most fairness adjustments.

As a consequence, in COMPAS, the limited standalone effectiveness of LD constrains the potential for complementarity. When LD produces negligible improvement, PS cannot systematically amplify fairness gains and instead induces metric-level variability. Hence, the degree of LD+PS synergy depends not only on heterogeneity level but also on the intrinsic sensitivity of each algorithmic family to fairness-aware local adjustments and sampling interventions.

5.1.6 Analysis of Combined Performance Metrics

To better analyze the interaction between bias-aware and accuracy-based metrics. Tables 5.4 and 5.5 present the top-performing methods by combined FAS (see Section 4.6) for Adult and COMPAS datasets at representative heterogeneity levels.

We observe the following:

Table 5.5: Top 5 methods by Combined FAS (COMPAS dataset, Dirichlet partitioning)

α	**Method**	**Acc**	**EOD**	**SPD**	**AOD**	**AccDiff**	**FAS**
High Heterogeneity ($\alpha = 0.1$)							
	FedCvg-Ratio(0.5)+LD+PS(0.5)	0.555	0.029	0.023	0.017	-0.058	**0.554**
	FedCvg-Ratio(0.5)+LD	0.554	0.028	0.025	0.018	-0.060	0.553
	FedCvg-Ratio(0.9)+LD	0.554	0.028	0.025	0.018	-0.060	0.553
	FedCvg-Ratio(0.9)+LD+PS(0.7)	0.554	0.027	0.022	0.017	-0.058	0.553
	FedCvg-Ratio(0.5)+PS(0.5)	0.554	0.028	0.022	0.017	-0.057	0.553
Moderate Heterogeneity ($\alpha = 0.5$)							
	FedCvg-Ratio(0.5)	0.522	0.032	0.027	0.019	-0.054	**0.519**
	FedCvg-Ratio(0.9)	0.522	0.032	0.027	0.019	-0.054	0.519
	FedCvg-Ratio(0.5)+LD	0.522	0.033	0.027	0.019	-0.054	0.519
	FedCvg-Ratio(0.9)+LD	0.522	0.033	0.027	0.019	-0.054	0.519
	FedCvg-Ratio(0.5)+PS(0.5)	0.521	0.031	0.028	0.019	-0.054	0.518
Low Heterogeneity ($\alpha = 5000$)							
	FedCvg-Ratio(0.5)	0.571	0.064	0.072	0.057	-0.052	**0.551**
	FedCvg-Ratio(0.9)	0.571	0.064	0.072	0.057	-0.052	0.551
	FedCvg(0.0001)	0.571	0.065	0.074	0.058	-0.051	0.550
	FedCvg(0.0001)+LD	0.570	0.064	0.074	0.058	-0.051	0.549
	FedCvg-Ratio(0.5)+LD	0.570	0.066	0.074	0.059	-0.049	0.549

- FedCvg-Ratio *dominates on both datasets.* Across most heterogeneity levels, FedCvg-Ratio variants consistently achieve the highest FAS on both Adult and COMPAS. For example, on Adult, FedCvg-Ratio+LD+PS reaches FAS = 0.784 at $\alpha = 5000$. On COMPAS, FedCvg-Ratio (without LD) achieves FAS = 0.551 at $\alpha = 5000$, confirming that ratio-based aggregation provides the strongest overall fairness–accuracy balance.

- *The effect of LD and PS is dataset-dependent.* The dataset-dependent effect of LD and PS, already observed in Sections 5.1.3 and 5.1.4, is highlighted also when considering FAS. Indeed, on Adult, all top-5 configurations, for each heterogeneity level, include LD. On COMPAS, as already discussed, LD yields minimal gains and the best-performing configuration at $\alpha = 5000$ does not include LD.

- *High heterogeneity leads to different optimal strategies.* When considering $\alpha = 0.1$, optimal strategies differ on the two datasets On Adult, FedCvg(0.0001)+LD+PS(0.5) achieves the best FAS (0.773), prioritizing strong accuracy (0.845) alongside substantial fairness gains. On COMPAS, FedCvg-Ratio(0.5)+LD+PS(0.5) achieves the best FAS (0.554). This divergence suggests that optimal techniques depend on dataset size and imbalance severity. We observe that in this regime, better results are anyway obtained when limiting the impact of coverage-based aggregation ($\alpha_{cov} = 0.0001$)

- *Moderate heterogeneity favors* FedCvg-Ratio. At $\alpha = 0.5$, for both datasets, FedCvg-Ratio variants lead to the highest FAS values. On Adult, FedCvg-Ratio(0.5)+LD+PS achieves the best FAS on Adult (0.778), showing that combining ratio-based aggregation, local debiasing, and parity sampling is most effective when data are moderately non-IID and LD operates reliably. As already observed in Section 5.1.3, LD is not effective on COMPAS.

- *Near-IID regimes benefit from PS on Adult.* At $\alpha = 5000$, FedCvg-Ratio variants still provides the best results for Adult (in combination with LD and PS). In particular, when considering LD with PS and parity sampling probablity equal to 0.5, FedCvg-Ratio improves FAS from 0.778 to 0.784, indicating that strategic client sampling remains beneficial even when distributions are nearly homogeneous. On COMPAS, FedCvg-Ratio and FedCvg variants have similar combined behavior and, as already observed, LD is not very effective.

- *Substantial FAS range differences across datasets.* Adult reaches FAS values in the range 0.770–0.784, whereas COMPAS ranges between 0.519 and 0.554, a gap of roughly 40%. This difference reflects structural dataset characteristics: COMPAS is smaller (6K vs. 49K samples), more imbalanced (81% vs. 67% male), and achieves lower overall accuracy (0.52–0.57 vs. 0.83–0.85). These factors impose intrinsic limits on attainable fairness–accuracy trade-offs.

5.2 Experiment 2: Coverage-Based Partitioning

The goal of the second experiment is to compare the reference algorithms under coverage-based data partitioning, which creates heterogeneity focused on representation disparity rather than label distribution. This partitioning strategy follows the methodology proposed in Brocchi's thesis [Bro23], where "good clients" satisfy a coverage constraint for the unprivileged group while "bad clients" have limited representation. Unlike Dirichlet partitioning, which creates stochastic heterogeneity across both labels and sensitive attributes, coverage-based partitioning provides explicit control over representation imbalance, allowing us to systematically evaluate how algorithms respond to varying degrees of minority group under-representation.

To this aim, we first provide information about client data distributions (Subsection 5.2.1); then, we present and discuss the obtained results with the goal of analyzing baseline aggregation algorithms (Subsection 5.2.2), the impact of local debiasing (Subsection 5.2.3), the impact of parity sampling (Subsection 5.2.4), the interaction between local debiasing and parity sampling (Subsection 5.2.5), and the top approaches when considering the FAS combined metric (Subsection 5.2.6).

For all the algorithm versions described in Section 4.4, we consider one configuration for each combination of parameter values pointed out in Table 4.2 and all the metrics listed in Section 4.6 but FAS, which will be discussed in Subsection 5.2.6.

Tables C.1–C.6 in Appendix C present the comprehensive performance comparison on the Adult dataset under coverage-based partitioning, while Tables D.1–D.6 in Appendix D present results for the COMPAS dataset. Results are organized by coverage value and partitioning method, with each table showing all metrics for one configuration. Following the FairFed paper methodology, we select the configuration corresponding to the best EOD value (closest to zero) across learning rates and report the mean across five random seeds. We observe that also in this case, COMPAS consistently exhibits negative AccDiff values.

5.2.1 Client Data Distribution

Before analyzing algorithm performance, we characterize the data heterogeneity created by coverage-based partitioning. Understanding client compositions is essential for interpreting fairness results, as representation imbalances at the client level directly impact the effectiveness of different fairness interventions.

Coverage-based partitioning explicitly controls representation disparity by designating a subset of clients as "good" (satisfying a coverage constraint) and the remainder as "bad" (poor minority representation). Given a coverage value c and tolerance τ, a "good client" must have at least $c \cdot (1 - \tau)$ unprivileged samples. We test the two variants presented in Subsection 4.3.2:

- **cov_same_size**: All clients have approximately the same total samples, varying only in group distribution.
- **cov_diff_size**: Both client sizes and group distributions vary, simulating realistic scenarios.

The number of clients varies by coverage: higher coverage produces fewer clients (e.g., Adult coverage equal to 4497 produces ≈7 clients), while lower coverage produces more clients (e.g., Adult coverage equal to 1999 produces ≈16 clients).

Table 5.6 summarizes the resulting client distributions (number of clients; percentage of total unprivileged samples that each coverage value represents; minimum, average, and maximum unprivileged samples per client) across the considered coverage values for both datasets and partitioning methods. Statistics are averaged over the 5 random seeds.

We observe the following:

Table 5.6: Client data distribution statistics under coverage-based partitioning (averaged over 5 seeds)

Dataset	Coverage	Partition Method	Num Clients	% of Total Unpriv	Min Unpriv	Mean Unpriv	Max Unpriv
Adult Dataset (Total unprivileged: 16,117 females)							
	4497	same_size	7	28%	450	4,664	4,950
	4497	diff_size	7	28%	380	4,664	5,200
	2570	same_size	13	16%	260	2,512	2,830
	2570	diff_size	13	16%	210	2,512	3,100
	1999	same_size	16	12%	200	2,041	2,200
	1999	diff_size	16	12%	180	2,041	2,400
COMPAS Dataset (Total unprivileged: 4,998)							
	1067	same_size	5	21%	107	1,000	1,170
	1067	diff_size	5	21%	90	1,000	1,250
	610	same_size	8	12%	61	625	670
	610	diff_size	8	12%	50	625	720
	474	same_size	11	9%	47	454	520
	474	diff_size	11	9%	40	454	550

- *Explicit control*: Unlike Dirichlet, where "good clients" vary by seed and α, coverage-based partitioning provides precise control. By construction (see Section 4.3.2) exactly 50% of clients are "good" (meeting coverage constraint) by design. This explicit structure enables more controlled evaluation of fairness interventions, particularly for algorithms like FedCvg that are designed to leverage coverage information.

- *Coverage as % of total*: Coverage values represent different percentages of total unprivileged samples. For example, in Adult, $cov = 4497$ corresponds to 28% of 16,117 total female samples, meaning each good client must have $\geq$28% of all females. This is a stringent constraint, resulting in only 7 clients total. In contrast, when $cov = 1999$ (12% of total females), we get 16 clients, each with fewer unprivileged samples but more opportunities for strategic selection via Parity Sampling.

- *Client count trade-off*: Lower coverage produces more clients but smaller per-client datasets. For example, in Adult, $cov = 1999$ produces 16 clients (avg 2,041 unprivileged each), while $cov = 4497$ produces 7 clients (avg 4,664 unprivileged each). This trade-off affects both training stability (fewer samples per client may reduce model quality) and fairness interventions (more clients provide more selection opportunities for Parity Sampling).

- *Partition method differences*: The diff_size method shows slightly wider ranges (e.g., in Adult, with $cov = 4497$ we obtain unprivileged samples in the 380-5,200 range while with the same_size method the range is 450-4,950), simulating more realistic

heterogeneity where both client sizes and compositions vary. However, mean values remain identical across methods, ensuring fair comparison.

We expect that the coverage value and client data distribution have an impact on the considered algorithms:

- FedAvg: Since it weights clients by dataset size, it will give equal influence to "good" and "bad" clients regardless of their minority representation. This propagates representation bias to the global model, particularly when bad clients dominate in number or size.
- FairFed: It adjusts weights based on fairness metrics computed on each client's data. Bad clients with poor minority representation may have undefined fairness metrics (e.g., no positive examples from one group), triggering the accuracy-gap fallback. The explicit 50% good/bad split should enable more consistent fairness metric computation compared to Dirichlet.
- FedCvg/FedCvg-Ratio: They explicitly account for representation imbalances by boosting clients with better minority representation. The explicit coverage structure should align well with their design: good clients will receive higher weights, directly addressing representation bias. Lower coverage values (more clients) may enable more fine-grained weight adjustments.
- Parity Sampling: With exactly 50% good clients, PS has clear targets to select. Lower coverage (more clients) provides more selection opportunities, potentially improving effectiveness compared to Dirichlet where good client identification is less explicit.
- Local Debiasing: Effectiveness depends on having sufficient samples for each combination of sensitive and target attributes. Bad clients with poor minority representation may have too few samples for effective reweighting, limiting LD's impact on those clients.

5.2.2 Comparison of Aggregation Algorithms

In this experiment, for each dataset, we compare the considered algorithm variants with respect to the coverage value for all the metrics under analysis. For FedCvg-Ratio, the two configurations corresponding to $\lambda = 0.5$ and $\lambda = 0.9$ yield nearly indistinguishable results across all reported metrics; therefore, for clarity, only the $\lambda = 0.5$ variant is shown.

Unless specified otherwise, the analysis focuses on the `diff_size` partitioning method, as it introduces a more realistic form of heterogeneity where client sizes vary. The trends

observed for `same_size` are generally consistent with those reported here, with specific differences noted in the Local Debiasing analysis (Subsection 5.2.3).

5.2.2.1 Adult Dataset

Figure 5.7 summarizes the baseline performance of the considered methods on the Adult dataset across different coverage values (the lower the coverage, the more clients).

Trend comparison. For all techniques, when increasing coverage (fewer clients), accuracy and accuracy difference improve while fairness metrics (EOD, SPD, AOD) degrade. From the plots, FedCvg behavior varies with both α_{cov} and coverage value. When $\alpha_{cov} = 0.001$, the impact of coverage becomes more pronounced, generally improving fairness at the cost of lower accuracy. FedCvg-Ratio shows more stable behavior across coverage values, achieving best values for all the metrics but SPD. The FairFed variants show similar overall trends across coverage levels, with differences reflecting the specific fairness objective they target.

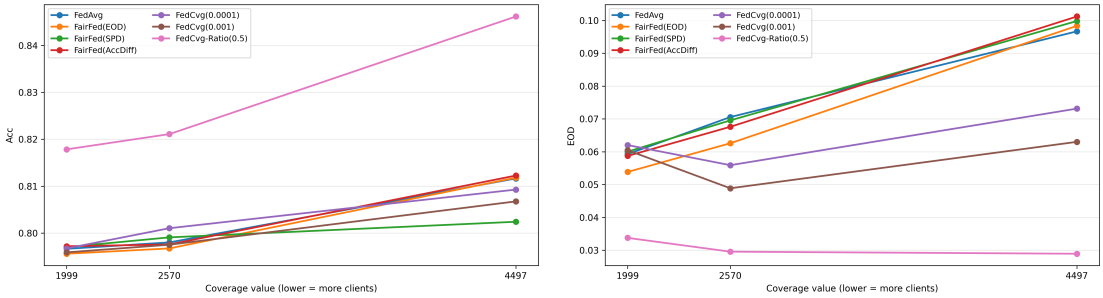
Accuracy. Predictive performance (Fig. 5.7(a)) improves for all techniques while decreasing the number of clients.

FedCvg-Ratio consistently achieves the highest accuracy across all coverage values: 0.818 at cov=1999, 0.821 at cov=2570, and 0.846 at cov=4497. FedAvg and FairFed variants achieve similar accuracy: approximately 0.797-0.799 at low/intermediate coverage, rising to 0.812 at high coverage. FedCvg variants show intermediate performance. The accuracy advantage of FedCvg-Ratio reaches a 4.2% improvement over FedAvg at cov=4497 (0.846 vs 0.812).These results indicate that representation-aware methods do not necessarily sacrifice accuracy, and can even enhance it in certain regimes by ensuring more robust model updates from representative clients.

Fairness metrics. Fairness trends reveal the strength of representation-based aggregation. In Fig. 5.7(b), FedCvg(0.001) and FedCvg-Ratio achieve markedly lower EOD values compared to other approaches across all coverage levels, with FedCvg-Ratio obtained the best results. For instance, with $cov = 1999$, FedAvg reports EOD= 0.059, while FedCvg-Ratio reduces it to 0.034, corresponding to an improvement of approximately 42% compared to FedAvg.

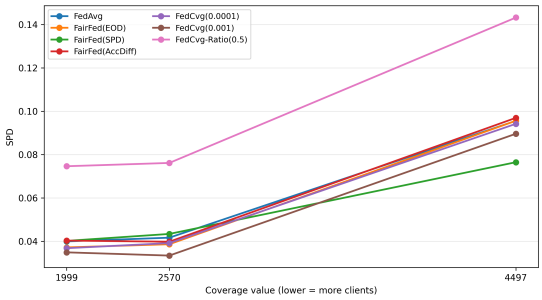
At $cov = 1999$, FedCvg-Ratio reduces also AOD to 0.029 compared to 0.035 of FedAvg (17% improvement). However, FedCvg-Ratio shows substantially worse SPD: 0.075 vs FedAvg's 0.040 (88% worse), highlighting a trade-off between equalized odds and demographic parity. FairFed variants perform nearly as well as FedAvg, with some slight variations in performance.

Accuracy difference. The AccDiff metric (Fig. 5.7(e)) improves while increasing the cov-

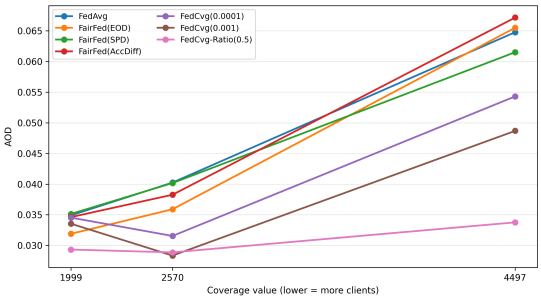


(a) Accuracy

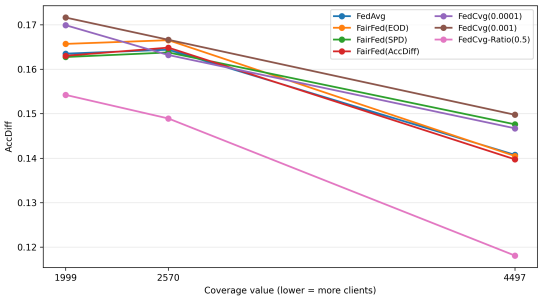
(b) EOD



(c) SPD



(d) AOD



(e) AccDiff

Figure 5.7: Baseline performance on the Adult dataset across coverage values (the lower the coverage, the more clients).

erage values. FedCvg-Ratio consistently achieves the lowest AccDiff across all coverage values: 0.154 at $cov = 1999$, 0.149 at $cov = 2570$, and 0.118 at $cov = 4497$. All the other approaches show higher and quite similar AccDiff values. The advantage of FedCvg-Ratio increases with coverage, indicating better inter-group accuracy parity with fewer, larger clients.

Conclusions. FedCvg-Ratio emerges as the strongest performer on Adult for all metrics except SPD. Coverage level creates a clear trade-off: lower coverage (more clients) produces better fairness metrics (EOD, SPD, AOD) but lower accuracy, while higher coverage (fewer clients) favors accuracy and AccDiff at the expense of fairness.

5.2.2.2 COMPAS Dataset

Figure 5.8 summarizes the baseline performance of the considered methods on the COMPAS dataset across different coverage values.

Trend comparison. On COMPAS, unlike Adult where FedCvg-Ratio dominated, here no single method excels across all metrics, reflecting the challenge of the extreme global imbalance (81% male). Additionally, the behavior is often mixed, with values for the intermediate coverage value often diverging from the main trends. We observe that the FedCvg-Ratio behavior is sometimes different from that of the other approaches; this happens for EOD and AOD. FairFed and FedAvg maintain more consistent performance across coverage levels for the all the considered metrics. FedCvg variants show clearer trends for accuracy and EOD, which improve while increasing the coverage value.

Accuracy. Accuracy increases substantially with coverage for all methods, with exceptions for the intermediate coverage value. Lower values are observed for representation-based methods, with FedCvg-Ratio achieving the worst results when increasing coverage. The accuracy ranking remains consistent across medium and high coverage levels: FairFed/FedAvg > FedCvg > FedCvg-Ratio. The extreme imbalance of the dataset penalizes representation-based methods, with FedCvg showing the poorest performance at low coverage (6% below FairFed).

Fairness metrics. Fairness behavior on COMPAS is complex and coverage-dependent. With the exception of FedCvg-Ratio on EOD and AOD, we observe that generally fairness metrics downgrade when increasing the coverage value, thus decreasing the number of clients. For EOD, FairFed(EOD) achieves the best value at low coverage (0.045 at $cov = 474$), substantially better than FedCvg-Ratio (0.070, 56% worse). However, EOD degrades for FairFed/FedAvg as coverage increases (0.045 $\rightarrow$ 0.064 for FairFed(EOD)), while FedCvg-based methods improve (0.070 $\rightarrow$ 0.061 for FedCvg-Ratio). At high coverage ($cov = 1067$), FairFed(AccDiff) achieves best EOD (0.053), followed by FedCvg-Ratio (0.061). For SPD, FedCvg variants excel at low/intermedi-

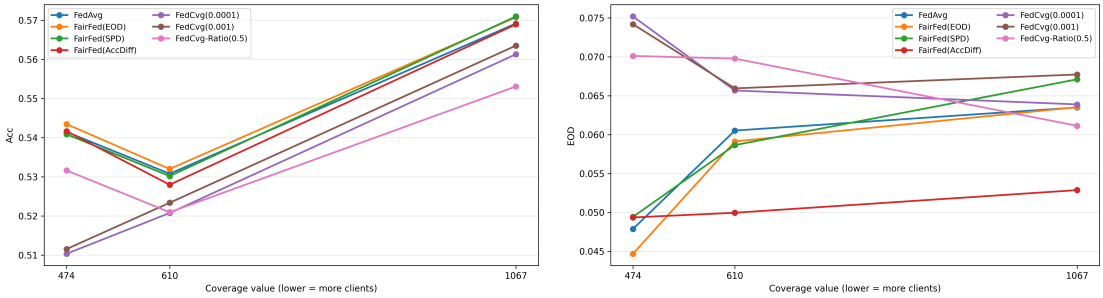
ate coverage (0.038 at $cov = 474$), while FedCvg-Ratio achieves best SPD at high coverage (0.043 vs FedAvg's 0.061). For AOD, FedCvg-Ratio achieves the best value at high coverage (0.031), significantly better than others (0.035-0.042). These mixed patterns confirm that representation-based aggregation struggles with extreme imbalance, where no single approach dominates across all fairness dimensions.

Accuracy difference. AccDiff improves (moves closer to zero) with higher coverage for all methods, indicating better inter-group parity with fewer, larger clients. At $cov = 474$, FedCvg(0.0001) achieves best AccDiff (-0.034), followed by FedCvg-Ratio (-0.035) and FedAvg/FairFed (-0.052 to -0.059). At $cov = 1067$, FedCvg(0.001) achieves best AccDiff (-0.010), followed by FedCvg(0.0001) (-0.012) and FedCvg-Ratio (-0.017), while FedAvg/FairFed range from -0.018 to -0.026. FedCvg variants consistently achieve the best inter-group accuracy parity, with improvements of 50-67% compared to FedAvg at high coverage. However, all values remain negative, indicating persistent higher accuracy for the privileged group (female), reflecting the structural challenge of COMPAS's extreme imbalance.

Conclusions. On COMPAS, no single method dominates across all metrics, reflecting the challenge of extreme imbalance. FairFed variants achieve the best accuracy and competitive fairness. FedCvg variants excel at AccDiff but suffer from poor accuracy at low coverage. FedCvg-Ratio provides a middle ground: often worst accuracy, good fairness metrics, especially for high coverage values. Coverage level significantly impacts performance: higher coverage improves accuracy (+5-6% from $cov = 474$ to $cov = 1067$) and AccDiff (improving by 50-67%), but fairness metrics show mixed trends (EOD worsens for FairFed/FedAvg, improves for FedCvg-based methods). The extreme sensitive attribute imbalance (81% male) fundamentally limits all methods, making algorithm choice highly dependent on which metrics are prioritized.

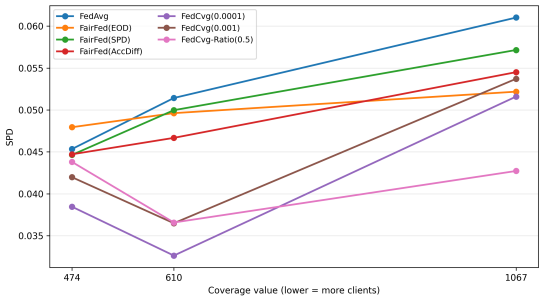
5.2.3 Impact of Local Debiasing

Figures 5.9 and 5.10 report the impact of Local Debiasing (LD) across different coverage values on the Adult and COMPAS datasets. Each bar in the histograms represents the variation $\Delta = (\mathcal{A} + LD) - \mathcal{A}$ for a specific metric, where $\mathcal{A}$ is one of the considered algorithms. Negative values indicate an improvement for fairness metrics (SPD, EOD, AOD, AccDiff), while positive values indicate an improvement in Accuracy. For FairFed, for SPD, EOD, and AccDiff we report only the variant trained on that metric, to simplify the analysis.

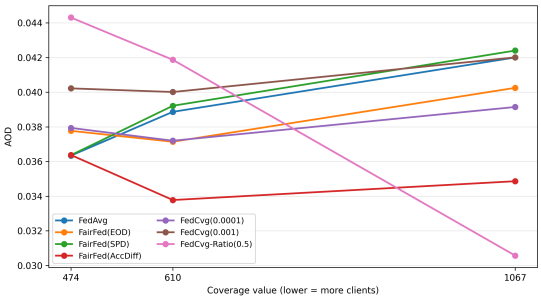


(a) Accuracy

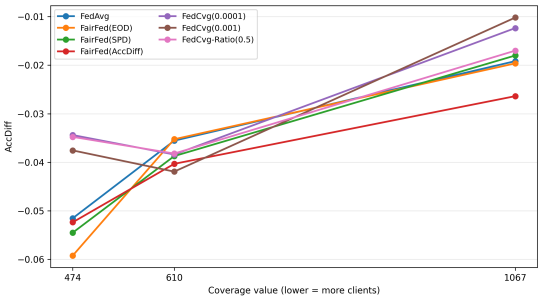
(b) EOD



(c) SPD



(d) AOD



(e) AccDiff

Figure 5.8: Baseline performance on the COMPAS dataset across coverage values (the lower the coverage, the more clients).

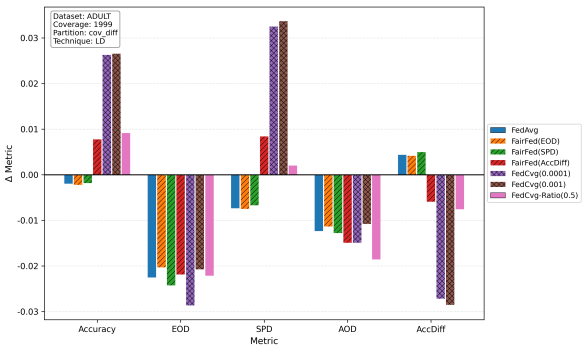
5.2.3.1 Adult Dataset

Trend comparison. Similar to the Dirichlet experiments, variations are generally modest (under 0.05 for most configurations). However, LD systematically improves fairness-related metrics across all coverage values. EOD, SPD, and AOD exhibit predominantly negative Δ values, indicating lower disparity after introducing LD. The impact on AccDiff and accuracy is less pronounced but generally positive at lower coverage values (more clients). The highest improvements are achieved by FedCvg(0.0001) and FedAvg, particularly at lower coverage values where more clients provide more opportunities for local reweighting. FedCvg(0.001) and FedCvg-Ratio show more modest improvements, likely because their representation-based aggregation already addresses much of the representation bias that LD targets.

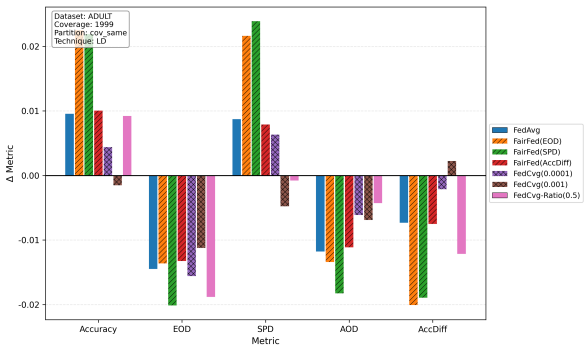
Coverage value impact. LD effectiveness varies with coverage level. At low coverage (cov = 1999, 16 clients), LD produces moderate fairness gains: FedAvg shows ΔEOD=-0.023, FedCvg(0.0001) shows ΔEOD=-0.029. At intermediate coverage (cov = 2570, 13 clients), effects are similar but slightly attenuated. Surprisingly, at high coverage (cov = 4497, 7 clients), LD produces the largest improvements: FedAvg achieves ΔEOD=-0.046, FairFed variants achieve ΔEOD=-0.045 to -0.052. This suggests that with fewer, larger clients, local reweighting becomes more effective, possibly because larger clients have sufficient samples for all (label, sensitive) combinations, enabling more robust reweighting. When considering accuracy and accuracy difference, the trend is different: best results are achieved with low coverage. Representation-aware techniques gives the best fairness results for all metrics except SPD and the best accuracy values for lower or intermediate coverage. This shows that with a higher number of clients, representation-aware are a valid option to mediate between fairness and accuracy.

Partitioning method. The diff_size partitioning method generally shows slightly larger LD effects compared to same_size, particularly for FedAvg and FedCvg(0.0001). This suggests that realistic client size variation creates additional opportunities for local reweighting to correct imbalances. However, the differences between partitioning methods are modest, indicating that LD's effectiveness is primarily determined by coverage value rather than client size homogeneity.

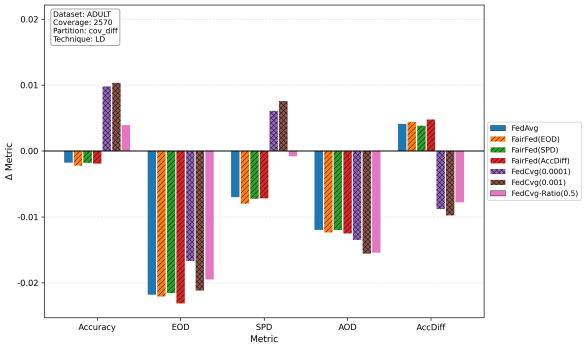
Conclusions. LD provides consistent fairness improvements under representation-based partitioning, particularly at lower coverage values where more clients enable more fine-grained local corrections. The effects are most pronounced for methods that do not already incorporate strong fairness mechanisms (FedAvg, FedCvg(0.0001)), while representation-based methods with stronger fairness constraints (FedCvg(0.001), FedCvg-Ratio) show more modest additional gains. The pattern mirrors Dirichlet results, confirming that LD is an effective complementary technique across different partitioning strategies.



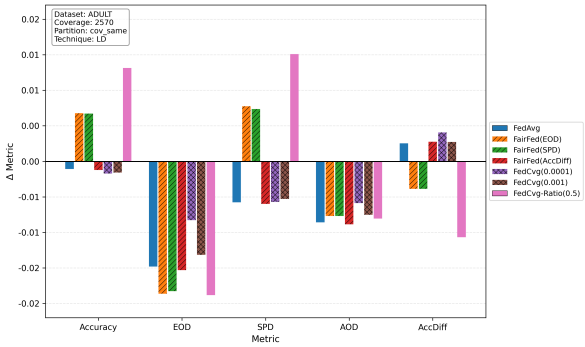
(a) Coverage = 1999, diff_size



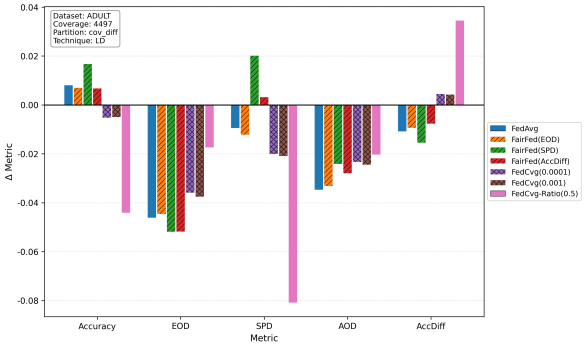
(b) Coverage = 1999, same_size



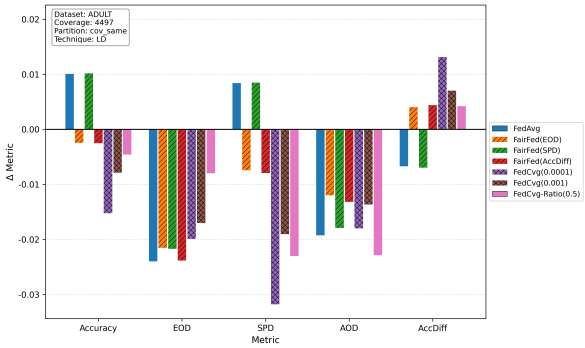
(c) Coverage = 2570, diff_size



(d) Coverage = 2570, same_size



(e) Coverage = 4497, diff_size



(f) Coverage = 4497, same_size

Figure 5.9: Impact of LD across coverage values on the Adult dataset.

5.2.3.2 COMPAS Dataset

Trend comparison. On COMPAS, LD effects are minimal and highly variable, mirroring the pattern observed under Dirichlet partitioning. Variations are mostly around zero, with no clear systematic trends across coverage values or partitioning methods. This confirms that LD's effectiveness is fundamentally limited by dataset characteristics rather than partitioning strategy.

Coverage value impact. Unlike Adult, coverage value has little impact on LD effectiveness on COMPAS. At all coverage levels, LD produces mixed results: some methods show slight improvements in specific metrics, while others show slight degradations. The extreme sensitive attribute imbalance (81% male) limits the number of minority samples available for reweighting, making local corrections ineffective regardless of how clients are partitioned.

Conclusions. LD does not provide systematic benefits on COMPAS under representation-based partitioning, consistent with Dirichlet results. The fundamental challenge is the extreme global imbalance, which leaves insufficient minority samples in most clients for effective local reweighting. This dataset-dependent limitation persists across different partitioning strategies, highlighting the importance of adequate minority representation for local debiasing techniques.

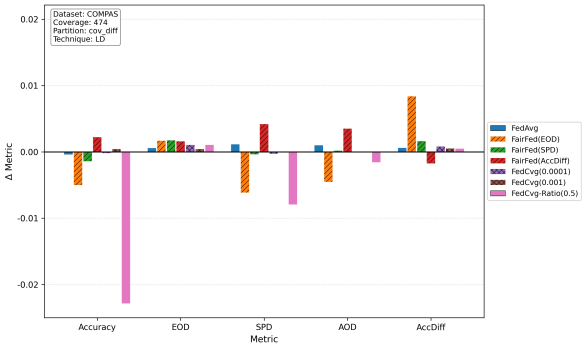
5.2.4 Impact of Parity Sampling

Figures 5.11 and 5.12 report the impact of Parity Sampling (PS) across different coverage values on the Adult and COMPAS datasets, showing results for both $p = 0.5$ and $p = 0.7$.

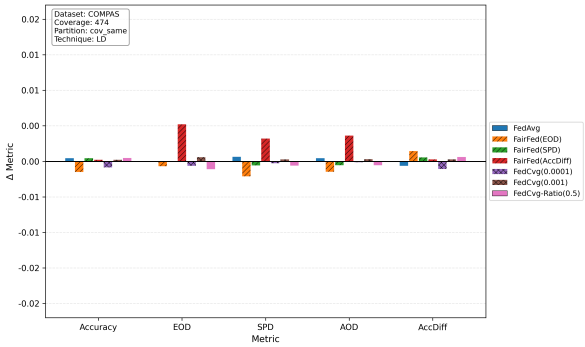
5.2.4.1 Adult Dataset

Trend comparison. The introduction of Parity Sampling under representation-based partitioning shows different patterns compared to Dirichlet. With the explicit 50% good/bad client structure, PS has clearer targets to select, potentially improving its effectiveness. However, the impact remains modest (variations up to 0.03) and varies by algorithm and coverage value.

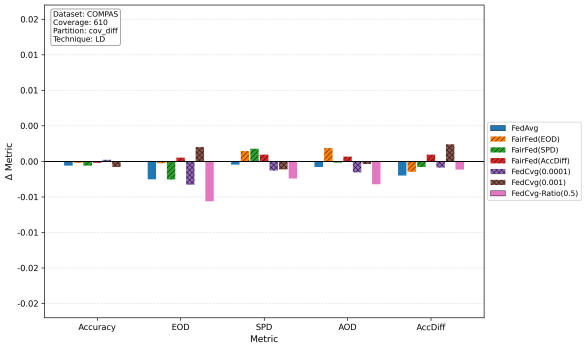
Overall, the FairFed family of algorithms tends to achieve the strongest improvements in fairness metrics when PS is applied. FedAvg also shows a mostly positive impact, although the magnitude of the improvement is generally smaller. Representation-based methods exhibit more variable behavior, sometimes improving and sometimes degrading fairness metrics. A possible explanation is that for FedCvg and FedCvg-Ratio we apply a coverage mechanism that already aims to mitigate representation bias, so applying



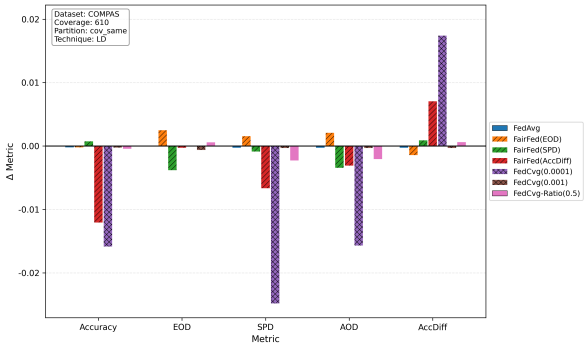
(a) Coverage = 474, diff_size



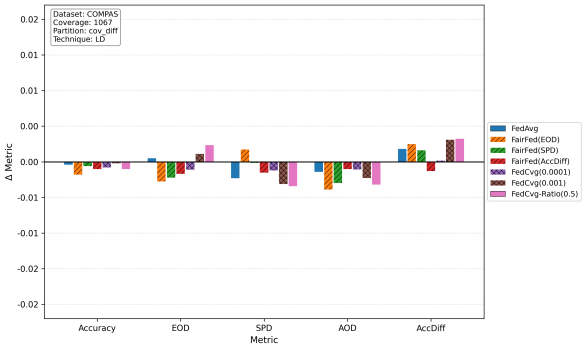
(b) Coverage = 474, same_size



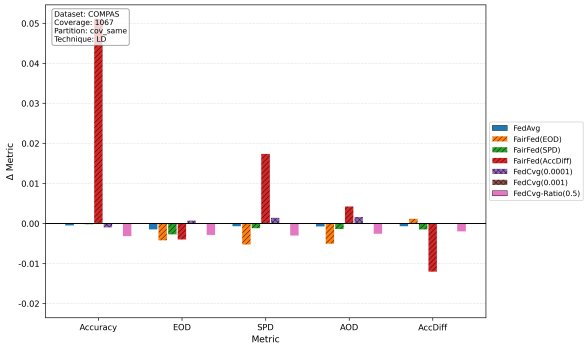
(c) Coverage = 610, diff_size



(d) Coverage = 610, same_size



(e) Coverage = 1067, diff_size



(f) Coverage = 1067, same_size

Figure 5.10: Impact of LD across coverage values on the COMPAS dataset.

PS—which targets a similar issue—provides limited additional benefit.

Coverage value impact. PS effectiveness varies significantly by algorithm and coverage level. For FedAvg and FairFed, PS shows increasing effectiveness with higher coverage: at *cov* = 1999, FedAvg achieves minimal ΔEOD=-0.003, while at cov=4497, it achieves ΔEOD=-0.007 to -0.010, and FairFed variants achieve ΔEOD=-0.010 to -0.015. This suggests that with fewer, larger clients, the quality difference between good and bad clients becomes more pronounced, making strategic selection more impactful.

However, FedCvg-Ratio shows anomalous behavior at low coverage: PS improves accuracy (+0.011) and EOD (-0.006) but substantially degrades SPD (+0.023), indicating interference between PS and representation-based aggregation. At higher coverage levels, FedCvg-Ratio shows minimal PS effects (ΔEOD=-0.001 to -0.004), suggesting that its internal weighting mechanism already performs implicit client selection, making explicit PS redundant or counterproductive.

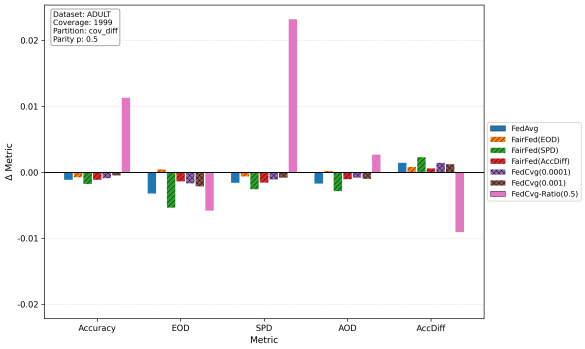
FedCvg does not show any notable improvement or degradation as the coverage value changes. Across the considered settings, both fairness metrics and accuracy remain relatively stable, indicating that the algorithm's performance is largely insensitive to variations in coverage in this scenario.

Parity probability. Varying p from 0.5 to 0.7 does not produce substantial changes in the effectiveness of PS across the considered algorithms. The overall trends remain largely consistent, with similar fairness variations observed at both values of p. For FedCvg-Ratio, the anomalous behavior at low coverage persists across both settings, suggesting that the effect is systematic rather than driven by the specific parity probability.

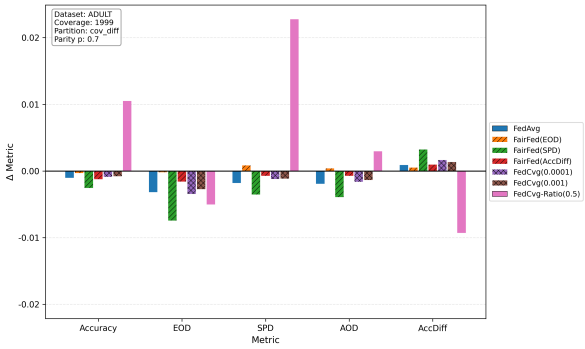
Conclusions. PS shows coverage-dependent effectiveness, with larger improvements at high coverage for methods without built-in fairness mechanisms (FedAvg, FairFed). However, PS interferes with representation-based aggregation (FedCvg-Ratio), particularly at low coverage where it creates trade-offs between different fairness metrics (improving EOD but degrading SPD). This suggests that PS is most beneficial for algorithms that do not already incorporate representation-aware weighting, and that combining multiple fairness interventions can produce unexpected interactions.

5.2.4.2 COMPAS Dataset

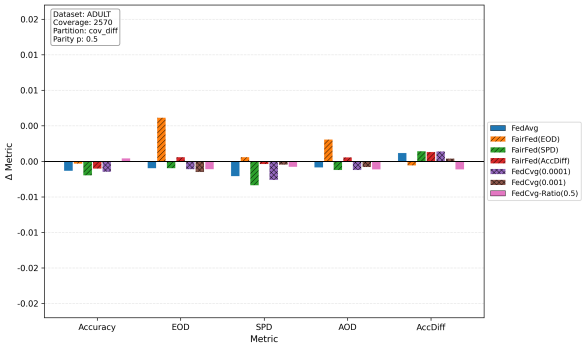
Trend comparison. On COMPAS, PS effects are even more limited and irregular than on Adult. Variations are mostly below 0.02 and even below 0.01 for intermediate and high coverage values, with no clear systematic patterns across coverage values or parity probabilities. The extreme sensitive attribute imbalance limits PS effectiveness, as even "good" clients have relatively poor minority representation. Considering fairness metrics, FedCvg-Ratio is the approach that consistently gets the highest benefits from PS for



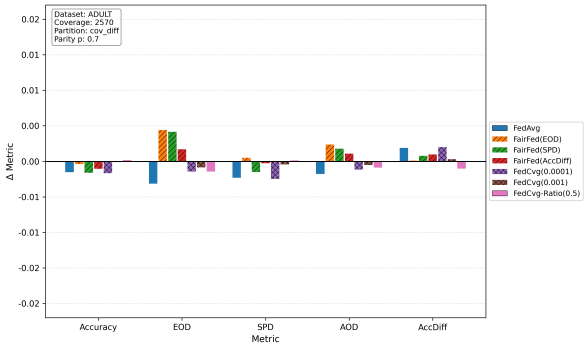
(a) Coverage = 1999, $p = 0.5$



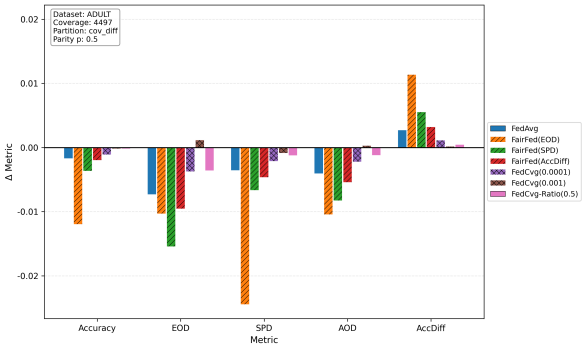
(b) Coverage = 1999, $p = 0.7$



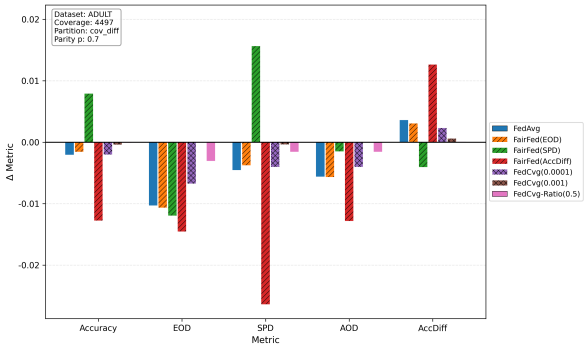
(c) Coverage = 2570, $p = 0.5$



(d) Coverage = 2570, $p = 0.7$



(e) Coverage = 4497, $p = 0.5$



(f) Coverage = 4497, $p = 0.7$

Figure 5.11: Impact of PS across coverage values on the Adult dataset (diff_size partitioning).

fairness metrics while the impact of PS on FairFed(AccDiff) is always negative.

Coverage value impact. Unlike Adult, coverage value has minimal impact on PS effectiveness on COMPAS. At all coverage levels, PS produces mixed and inconsistent results. Some methods show slight improvements in specific metrics at specific coverage values, but these patterns do not generalize across configurations.

Parity probability. Overall, parity sampling does not lead to consistent improvements in fairness metrics across the considered algorithms in this setting. A likely explanation is the smaller size of the dataset compared to Adult, which reduces the variability among clients and therefore limits the potential benefits of strategic sampling. Nevertheless, at higher coverage values—corresponding to fewer but larger clients—the FairFed variants, particularly FairFed (EOD), achieve the most noticeable fairness improvements.

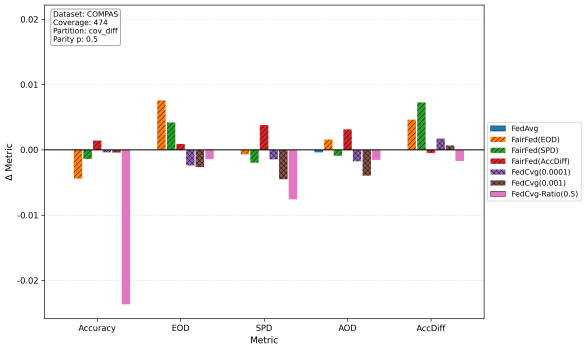
Conclusions. PS does not provide systematic benefits on COMPAS under representation-based partitioning, mirroring the pattern observed under Dirichlet. The fundamental challenge is the extreme global imbalance, which limits the quality of even "good" clients. Strategic client selection cannot overcome this structural limitation, regardless of partitioning strategy.

5.2.5 Interaction of Local Debiasing with Parity Sampling

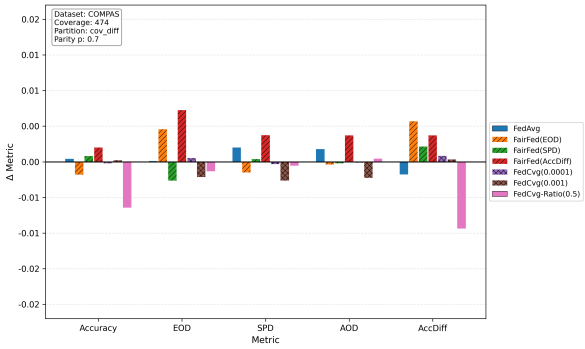
After having discussed the impact of LD and PS separately, we now discuss their joint effect under coverage-based partitioning. To this aim, Table 5.7 presents a comprehensive analysis of EOD values at representative coverage levels for both Adult and COMPAS datasets, including all algorithm variants. We consider EOD as reference metric since the results reported in the tables in the appendix corresponds to the configurations leading to the best EOD values among all the considered learning rates.

Adult dataset. On Adult, the LD+PS combination shows stronger synergy under coverage-based partitioning, particularly for representation-aware methods. At $cov = 1999$, FedCvg-Ratio(0.5) achieves remarkable results: baseline EOD= 0.034, LD alone reduces it to 0.012 (65% improvement), and LD+PS further reduces it to 0.009 (74% improvement over baseline, 25% improvement over LD alone). This represents the best EOD value achieved across all experiments on Adult. The synergy is also visible for FedAvg, which moves from 0.059 (baseline) to 0.037 (LD) to 0.035 (LD+PS), showing consistent but more modest complementarity.

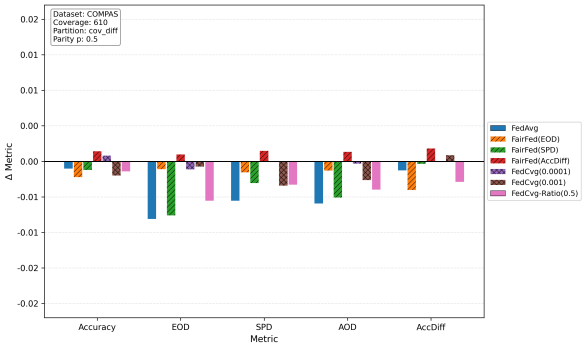
The explicit 50% good/bad client structure in coverage-based partitioning enables more effective PS compared to Dirichlet. When LD has already improved local fairness within clients, PS can strategically select the best-performing clients, amplifying the overall fairness gain. This synergy is strongest for FedCvg-Ratio, where representation-based ag-



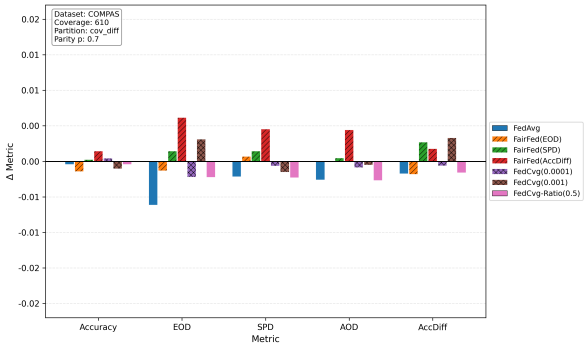
(a) Coverage = 474, $p = 0.5$



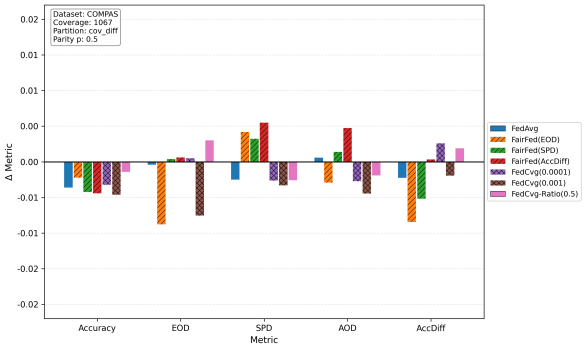
(b) Coverage = 474, $p = 0.7$



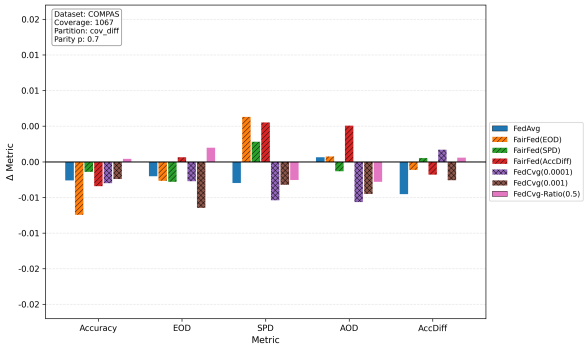
(c) Coverage = 610, $p = 0.5$



(d) Coverage = 610, $p = 0.7$



(e) Coverage = 1067, $p = 0.5$



(f) Coverage = 1067, $p = 0.7$

Figure 5.12: Impact of PS across coverage values on the COMPAS dataset (diff_size partitioning).

Table 5.7: Effect of Local Debiasing and Parity Sampling combinations under coverage-based partitioning (EOD values, diff_size method)

	Adult (cov=1999)				COMPAS (cov=474)			
Algorithm	**Baseline**	**PS**	**LD**	**LD+PS**	**Baseline**	**PS**	**LD**	**LD+PS**
FedAvg	0.059	0.056	0.037	0.035	0.048	0.048	0.048	0.047
FedCvg(0.0001)	0.062	0.060	0.033	0.034	0.075	0.073	0.076	0.076
FedCvg(0.001)	0.060	0.058	0.040	0.036	0.074	0.072	0.075	0.075
FedCvg-Ratio(0.5)	0.034	0.028	0.012	0.009	0.070	0.069	0.071	0.071
FedCvg-Ratio(0.9)	0.034	0.032	0.012	0.011	0.070	0.071	0.071	0.071
FairFed(EOD)	0.054	0.054	0.034	0.036	0.045	0.052	0.046	0.045
FairFed(AccDiff)	0.059	0.057	0.037	0.035	0.049	0.050	0.051	0.051
FairFed(SPD)	0.060	0.055	0.036	0.032	0.049	0.054	0.051	0.047

gregation, local reweighting, and strategic selection work together harmoniously. Notably, while PS alone can interfere with FedCvg-Ratio's representation-based weighting (degrading SPD), combining PS with LD mitigates this interference, suggesting that LD's local corrections enable more effective strategic selection.

For FairFed variants, the synergy is more variable. FairFed(SPD) shows clear complementarity (0.060 $\rightarrow$ 0.036 $\rightarrow$ 0.032), while FairFed(EOD) shows interference: LD achieves EOD= 0.034, but adding PS degrades it to 0.036. This suggests that when fairness objectives are already embedded in the aggregation mechanism, PS can interfere with the optimization, and the benefit of combining techniques depends on the specific fairness metric being targeted.

COMPAS dataset. On COMPAS, the LD+PS combination provides minimal benefits, mirroring the pattern observed under Dirichlet. At $cov = 474$, most methods show negligible differences between LD and LD+PS configurations. FedAvg shows a tiny improvement (0.048 $\rightarrow$ 0.047), while representation-based methods show no change (FedCvg-Ratio: 0.071 $\rightarrow$ 0.071). FairFed(EOD) shows a slight improvement (0.046 $\rightarrow$ 0.045), but the magnitude is negligible. The extreme sensitive attribute imbalance (81% male) fundamentally limits both LD and PS effectiveness, leaving little room for synergy between the two techniques.

Conclusions. The LD+PS combination is most effective under coverage-based partitioning on Adult, particularly for FedCvg-Ratio, achieving the best fairness results across all experiments. The explicit good/bad client structure enables stronger synergy compared to Dirichlet, where good client identification is less clear. However, this synergy is both dataset-dependent and algorithm-dependent: on COMPAS, extreme global imbalance limits both techniques individually and in combination; on Adult, FairFed(EOD)

Table 5.8: Top 5 methods by Combined FAS (Adult dataset, coverage-based partitioning, diff_size method)

Coverage	Method	Acc	EOD	SPD	AOD	AccDiff	FAS
Low Coverage (1999, 16 clients)							
	FedCvg-Ratio(0.9)+LD+PS(0.7)	0.828	0.008	0.080	0.013	0.146	**0.777**
	FedCvg-Ratio(0.5)+LD	0.827	0.012	0.077	0.011	0.147	0.776
	FedCvg-Ratio(0.9)+LD	0.827	0.012	0.077	0.011	0.147	0.776
	FedCvg(0.0001)+LD	0.823	0.033	0.069	0.020	0.143	0.768
	FedCvg(0.0001)+LD+PS(0.5)	0.822	0.034	0.068	0.020	0.144	0.767
Intermediate Coverage (2570, 13 clients)							
	FedCvg-Ratio(0.5)+LD	0.825	0.010	0.075	0.013	0.141	**0.776**
	FedCvg-Ratio(0.9)+LD	0.825	0.010	0.075	0.013	0.141	0.776
	FedCvg-Ratio(0.9)+LD+PS(0.5)	0.824	0.011	0.073	0.012	0.142	0.775
	FedCvg-Ratio(0.5)+LD+PS(0.5)	0.824	0.010	0.072	0.013	0.142	0.775
	FedCvg-Ratio(0.5)+LD+PS(0.7)	0.824	0.014	0.073	0.011	0.142	0.775
High Coverage (4497, 7 clients)							
	FedCvg-Ratio(0.9)+PS(0.7)	0.846	0.025	0.142	0.032	0.118	**0.779**
	FedCvg-Ratio(0.5)+PS(0.7)	0.846	0.026	0.142	0.032	0.118	0.779
	FedCvg-Ratio(0.5)+PS(0.5)	0.846	0.025	0.142	0.033	0.119	0.779
	FFedCvg-Ratio(0.9)+PS(0.5)	0.846	0.028	0.143	0.033	0.118	0.778
	FedCvg-Ratio(0.5)	0.846	0.029	0.143	0.034	0.118	0.778

shows interference between LD and PS rather than synergy. The results confirm that technique combinations are most beneficial when individual techniques are already effective and compatible, and that dataset characteristics (size, balance) are more important than partitioning strategy for determining technique effectiveness.

5.2.6 Analysis of Combined Performance Metrics

To better analyze the interaction between bias-aware and accuracy-based metrics, Tables 5.8 and 5.9 present the top-performing methods by combined FAS (see Section 4.6) for Adult and COMPAS datasets at each coverage level using diff_size partitioning.

We observe the following:

- FedCvg-Ratio *dominates on Adult across all coverage levels*: FedCvg-Ratio variants occupy all top-5 positions at all three coverage levels, demonstrating consistent excellence. Best FAS ranges from 0.776-0.779 across coverage levels, comparable to the best Dirichlet results (0.770-0.784). This dominance is even stronger than under Dirichlet partitioning, showing that FedCvg-Ratio is particularly well-suited to explicit coverage constraints.

Table 5.9: Top 5 methods by Combined FAS (COMPAS dataset, coverage-based partitioning, diff_size method)

Coverage	Method	Acc	EOD	SPD	AOD	AccDiff	FAS
Low Coverage (474, 11 clients)							
	FairFed(EOD)	0.543	0.045	0.048	0.038	-0.059	**0.534**
	FairFed(EOD)+LD+PS(0.7)	0.543	0.046	0.047	0.038	-0.057	0.533
	FairFed(AccDiff)+LD	0.544	0.051	0.049	0.040	-0.054	0.532
	FairFed(AccDiff)+LD+PS(0.5)	0.544	0.051	0.050	0.041	-0.055	0.532
	FairFed(SPD)+LD+PS(0.5)	0.542	0.047	0.046	0.037	-0.052	0.532
Intermediate Coverage (610, 8 clients)							
	FedAvg+LD+PS(0.5)	0.530	0.050	0.045	0.031	-0.038	**0.518**
	FedAvg+PS(0.5)	0.530	0.052	0.046	0.033	-0.037	0.517
	FairFed(EOD)	0.532	0.059	0.050	0.037	-0.035	0.517
	FairFed(SPD)+LD+PS(0.5)	0.530	0.052	0.046	0.033	-0.037	0.517
	FairFed(EOD)+LD	0.532	0.059	0.051	0.039	-0.037	0.517
High Coverage (1067, 5 clients)							
	FairFed(AccDiff)	0.569	0.053	0.055	0.035	-0.026	**0.552**
	FairFed(AccDiff)+LD	0.568	0.051	0.053	0.034	-0.028	0.552
	FairFed(EOD)+PS(0.5)	0.569	0.055	0.056	0.037	-0.028	0.551
	FairFed(EOD)	0.571	0.064	0.052	0.040	-0.020	0.551
	FairFed(EOD)+LD+PS(0.7)	0.569	0.059	0.056	0.038	-0.025	0.551

- *Coverage value creates accuracy-fairness trade-off*: At low coverage (1999, 16 clients), best EOD=0.008 with FAS=0.777 and accuracy=0.828. At high coverage (4497, 7 clients), best EOD=0.025 with FAS=0.779 and accuracy=0.846. Lower coverage produces substantially better fairness (68% improvement) but lower accuracy (2% reduction), resulting in comparable FAS. This trade-off reflects the tension between having more clients for strategic selection (better fairness) versus having larger clients for stable training (better accuracy).

- *Technique combinations vary by coverage*: At low and intermediate coverage, the best methods use LD or LD+PS, leveraging local reweighting to improve fairness. At high coverage, the best methods use PS only (without LD), suggesting that when clients are fewer but larger, local reweighting becomes less critical as clients already have more balanced internal distributions. This pattern highlights the importance of matching techniques to data characteristics.

- FairFed *dominates on COMPAS*: Unlike Adult, FairFed variants occupy all or most top-5 positions across all COMPAS coverage levels: 5/5 at low coverage (474), 3/5 at intermediate coverage (610, with FedAvg occupying 2/5), and 5/5 at high coverage (1067). FedCvg-Ratio, which excels on Adult, does not appear in any COMPAS top-5. This dramatic reversal confirms that representation-based aggre-

gation struggles with extreme imbalance (81% male), where explicit coverage constraints cannot overcome fundamental data limitations. FairFed's explicit fairness optimization is more effective in this challenging regime.

- *Coverage level impacts algorithm choice on COMPAS*: At low coverage (474, 11 clients), FairFed(EOD) achieves best FAS=0.534 with accuracy=0.543. At high coverage (1067, 5 clients), FairFed(AccDiff) achieves best FAS=0.552 with accuracy=0.569. Higher coverage produces fewer, larger clients, leading to higher accuracy (5% improvement) and better FAS despite similar fairness. This suggests that on highly imbalanced datasets, accuracy dominates FAS, and having fewer, larger clients benefits overall performance.

- *Comparison with Dirichlet*: Adult achieves FAS=0.776-0.779 under coverage-based vs 0.770-0.784 under Dirichlet (comparable). COMPAS achieves FAS=0.517-0.552 under coverage-based vs 0.519-0.554 under Dirichlet (nearly identical). This consistency across partitioning strategies validates that dataset characteristics (size, balance) are more important than partitioning method for overall performance. However, coverage-based partitioning achieves better absolute fairness on Adult (best EOD: 0.008 vs Dirichlet: 0.015), demonstrating that explicit representation control enables superior fairness when accuracy can be slightly sacrificed.

5.3 Summary

Based on the experimental results across both Dirichlet and coverage-based partitioning, we provide algorithm family recommendations for different deployment scenarios. Table 5.10 summarizes these recommendations, organized by scenario characteristics.

The table presents five representative scenarios encountered in federated learning deployments. For each scenario, we identify the recommended algorithm family (column 2), describe its key performance characteristics (column 3), and provide references to the specific experimental results supporting the recommendation (column 4). The scenarios span different heterogeneity levels (high, moderate, near-IID), optimization priorities (fairness-focused, balanced, accuracy-focused), and operational constraints (availability of local debiasing, client-side intervention capabilities).

From a practical standpoint, the results suggest differentiated guidance depending on the optimization objective:

- *Balanced objective.* FedAvg+LD provides a simple yet effective baseline, offering solid fairness gains without substantial architectural modifications. This approach

Table 5.10: Algorithm family recommendations by scenario

Scenario	**Recommended Family**	**Key Characteristics**	**Reference**
High heterogeneity, prioritize fairness	Representation-based + LD	Best fairness (low EOD/AOD), moderate accuracy	Figure 5.1
Moderate heterogeneity, balanced goals	Representation-based + LD	High FAS, balanced accuracy-fairness	Figures 5.1, 5.2
Near-IID, maximize both	Representation-based + LD + PS	Highest FAS, excellent accuracy and fairness	Tables 5.4, 5.5
Simple baseline	FedAvg + LD	Good fairness improvement, high accuracy	Figure 5.3, Table 5.7
No local debiasing available	FedCvg-Ratio	Moderate fairness without client-side intervention	Table 5.3, 5.7

Note: LD = Local Debiasing, PS = Parity Sampling. Representation-based refers to FedCvg/FedCvg-Ratio *family; fairness-aware refers to the* FairFed *family.*

achieves good accuracy-fairness trade-offs across heterogeneity levels while requiring minimal changes to standard federated learning infrastructure.

- *Fairness-priority objective.* FedCvg or FedCvg-Ratio combined with LD yield substantial reductions in EOD (often 50–70% relative improvement compared to baseline), at the cost of modest accuracy reductions (approximately 2–5%). These methods are particularly effective when fairness constraints are strict and some accuracy sacrifice is acceptable.

- *Accuracy-priority objective.* Under low heterogeneity (near-IID conditions), FedCvg-Ratio+LD+PS achieves near-optimal accuracy while preserving competitive fairness, making it a strong candidate when predictive performance remains the primary concern while still maintaining acceptable fairness levels.

The recommendations reflect the following key findings from our experiments:

- Performance-driven methods like FedAvg Achieve the highest predictive accuracy across most settings, but produce larger fairness disparities because the aggregation process ignores the distribution of sensitive groups across clients.

- Fairness-aware aggregation methods, like the FairFed variants, provide more balanced fairness outcomes than FedAvg while maintaining competitive accuracy. Improvements are generally moderate but stable, especially in highly non-IID settings.

Their effectiveness depends on the dataset: it remains competitive on Adult but becomes the most reliable strategy on COMPAS, where extreme imbalance limits the effectiveness of representation-based approaches.

- Representation-based methods (FedCvg/FedCvg-Ratio) consistently achieve the best balance between accuracy and fairness across heterogeneity levels, making them the preferred choice for most scenarios. They shows clear advantages on the Adult dataset, while differences between methods are smaller and more metric-dependent on COMPAS, which benefits more from fairness-aware optimization.

- Local debiasing is essential for all algorithm families, providing substantial fairness improvements with minimal accuracy cost; this makes FedAvg+LD a strong simple baseline when advanced methods are not feasible.

- Parity sampling mostly affects methods without built-in fairness mechanisms. It provides additional benefits primarily in near-IID settings where multiple well-represented clients enable effective strategic selection, while offering limited gains under high heterogeneity. For scenarios without client-side intervention capabilities (no LD), coverage-based methods with high sensitivity parameters or fairness-aware methods with server-side parity sampling offer the best alternatives, though with reduced effectiveness compared to LD-enabled configurations.

- Lower coverage (more clients with smaller datasets) tends to improve fairness by increasing diversity in client representations, while higher coverage (fewer but larger clients) improves accuracy due to more stable local training.

Overall, FedCvg-Ratio achieves robust performance across metrics and heterogeneity levels, making it a strong default choice for representation-aware federated learning.

Chapter 6

Conclusions

This thesis addressed the challenge of ensuring fairness in federated learning systems, where data heterogeneity and representation bias can lead to discriminatory outcomes for under-represented groups. We conducted a comprehensive comparative evaluation of fairness-aware federated learning algorithms, including the baseline FedAvg, the fairness-aware aggregation FairFed, the coverage-based FedCvg, and our novel ratio-based FedCvg-Ratio. All techniques were combined with local debiasing and parity sampling. Through extensive experiments on Adult Income and COMPAS datasets under various heterogeneity conditions, we demonstrated fairness improvements of 69–86% over baseline, often without sacrificing accuracy.

We conducted a comprehensive comparative evaluation of fairness-aware federated learning algorithms, including the baseline algorithm FedAvg, the fairness-aware aggregation approach FairFed, the coverage-based method FedCvg, and a new ratio-based FedCvg-Ratio approach, defined in the context of this work. All the techniques have been also combined with local debiasing at the client side and parity sampling for client selection. Through extensive experiments on the Adult Income and COMPAS datasets under various heterogeneity conditions, we demonstrated that fairness improvements of 69–86% over the baseline are achievable, often without sacrificing accuracy.

In the following, we summarize the main achieved contributions and the limitations of our work, we highlight future work directions, and finally we present some final concluding remarks.

6.1 Summary of the Contributions

After an accurate background analysis, we surveyed existing approaches to fairness enhancement in federated learning, identifying two main categories of bias: algorithmic bias (unfair model predictions) and representation bias (unequal group representation across clients). This analysis revealed gaps in the field, particularly the lack of systematic comparisons across different fairness interventions and the need for adaptive algorithms that do not require manual parameter tuning.

Our work makes four main contributions to fair federated learning. First, we provide a systematic comparison of fairness-aware algorithms across multiple dimensions: heterogeneity levels (Dirichlet α from 0.1 to 5000), partitioning strategies (Dirichlet and coverage-based), and datasets with different characteristics (Adult: 48K samples, 33% female; COMPAS: 6K samples, 19% female). Second, we introduce FedCvg-Ratio, a novel algorithm that replaces FedCvg's fixed coverage threshold with dynamic, ratio-based weighting that adapts automatically through exponential moving average smoothing. Third, we demonstrate that server-side and client-side fairness interventions are complementary: local debiasing provides substantial improvements (37–50% EOD reduction) when combined with any server-side algorithm. Fourth, we provide evidence-based algorithm recommendations for common deployment scenarios, enabling practitioners to select appropriate approaches based on data characteristics and fairness requirements.

The experimental results challenge several common assumptions about fairness in machine learning. Most notably, the fairness-accuracy trade-off is not universal in federated settings. FedCvg-Ratio combined with local debiasing can achieve substantial fairness improvements while simultaneously improving accuracy, because representation-based methods address representation bias that causes underfitting on minority groups. However, algorithm effectiveness depends critically on data characteristics: on balanced datasets like Adult, both FairFed and FedCvg-Ratio excel, while on highly imbalanced datasets like COMPAS, fairness-aware aggregation becomes more competitive. Heterogeneity level also matters: under high heterogeneity, representation-based methods show strong advantages, while at near-IID conditions, combining multiple techniques achieves best performance.

Local debiasing emerges as the most reliable fairness intervention, providing consistent improvements with minimal accuracy impact across all algorithms and scenarios but making assumptions on the used client-side learning approaches. Parity sampling shows more limited and context-dependent effects: it provides minimal benefit when combined with FedCvg (which already addresses representation bias), but shows noticeable improvements with FedAvg, particularly in near-IID configurations. Our novel FedCvg-Ratio eliminates manual coverage threshold tuning through ratio-based weighting and achieves robust performance across all metrics and heterogeneity levels, making it a strong default choice for practitioners.

6.2 Limitations

Our work has several limitations that should be considered when interpreting the results and can be summarized as follows:

- *Dataset limitations.* The experimental evaluation is limited to two datasets (Adult and COMPAS) with binary sensitive attributes and binary classification tasks. Real-world applications may involve multiple sensitive attributes or intersectional fairness concerns, where fairness must be ensured not just for individual groups but for their intersections. Our evaluation also focuses on tabular data and does not cover image, text, or other data modalities that are common in federated learning deployments.
- *Simulation environment.* All experiments were conducted in simulation where clients are processes on a single machine, not real distributed systems. Real deployments involve latency, bandwidth constraints, and client dropout that may affect algorithm performance. Additionally, clients in our experiments have identical computational resources, while real systems exhibit system heterogeneity with varying compute capabilities.
- *Only binary classification.* The algorithms we evaluate are designed for binary classification with known sensitive attributes available at training time. Extension to multi-class classification requires further work, as fairness definitions become more complex with multiple classes. Our data partitioning is also static, and dynamic client participation patterns where clients join and leave the federation over time are not evaluated.
- *Privacy issues not addressed.* Privacy considerations represent an important limitation. FairFed requires clients to share fairness metrics (TP, FP, TN, FN per group), while FedCvg and FedCvg-Ratio require clients to share group counts. This information sharing may reveal sensitive information about client data distributions and potentially enable inference attacks. We do not provide differential privacy guarantees or formal privacy analysis, and the privacy-fairness-accuracy trade-off remains an open question in our work.

6.3 Future Work

Future extensions of our work concern two main directions: the performed experimental evaluation and privacy issues. The main topics of interest for future work can be summarized as follows.

Experimental Analysis Extensions. Several directions can extend and deepen our experimental evaluation:

- *Hybrid partitioning method.* While our current experiments adopt two separate partitioning strategies—Dirichlet-based partitioning inspired by FairFed to generate heterogeneous distributions of the sensitive attribute across clients, and coverage-based partitioning to control representation bias—a future research direction would be to design a hybrid partitioning method that combines these two aspects.

  One possible approach is a ratio-guided Dirichlet partitioning scheme. In this setting, client datasets would still be generated through a Dirichlet sampling process, but the sampling would be guided by constraints on the representation ratio of the sensitive attribute, similarly to the idea underlying the FedCvg-Ratio algorithm.

  Such a strategy would allow the generation of federated scenarios where both distributional heterogeneity and representation imbalance coexist in a controlled manner. This would better reflect real-world federated environments, where clients typically differ both in their data distribution and in the representation of protected groups, and would enable a more systematic analysis of how fairness-aware algorithms behave under combined sources of bias.

- *Broader evaluation*: We aim to extend the performed experimental evaluation by considering other datasets, including datasets with natural partitions like the ACSIncome Dataset [Ope22], to understand the impact of FedCvg-Ratio and our fairness interventions over real data distributions. The experimental evaluation should also be extended to include multi-class classification tasks, datasets with multiple sensitive attributes, and data from different domains such as healthcare or finance where fairness requirements may differ.

- *Comparison with other methods.* The experimental evaluation should be extended to compare our approach to other discrimination-aware methods by implementing all solutions within the same environment, ensuring fair comparison under identical conditions.

- *Intersectional fairness.* We plan to broaden the proposed solutions to consider additional constraints for identifying representation bias, including coverage constraints defined over different sensitive attributes, following an intersectional approach, and multiple coverage constraints.

Privacy-Preserving Fairness. Privacy considerations represent a critical direction for future work, as the tension between fairness and privacy in federated learning remains largely unexplored:

- *Secure aggregation integration.* As discussed in Chapter 2, our current algorithms require clients to share fairness-related statistics such as group counts and confusion matrix elements, which may leak sensitive information about client data distributions. Recent work on secure aggregation techniques, such as per-element SecAgg [SKTH25], provides cryptographic mechanisms to protect against data reconstruction attacks by ensuring that aggregated values are revealed only when sufficient clients contribute. This approach has already been used in the initial design of FedAvg [MMR+17] and FairFed [EYH+23]. Integrating such techniques into fairness-aware federated learning represents a promising direction, though it requires careful design to handle the fine-grained statistics needed for fairness interventions. The computational overhead and communication costs of secure aggregation protocols also need evaluation in the context of representation-based algorithms.

- *Differential privacy integration.* Beyond secure aggregation, the integration of differential privacy into fairness-aware algorithms presents an additional challenge, as adding noise to protect privacy may degrade both accuracy and fairness. Understanding the three-way trade-off between privacy, fairness, and accuracy requires both theoretical analysis and empirical evaluation. Investigating how to add differential privacy guarantees to fairness-aware algorithms while maintaining acceptable performance is an important open problem.

- *Federated analytics.* Developing privacy-preserving methods for computing global fairness metrics without centralizing sensitive data could reduce privacy risks while maintaining the ability to monitor and enforce fairness requirements. Federated analytics approaches that compute global fairness metrics in a distributed manner, potentially leveraging secure aggregation, represent a promising direction.

6.4 Final Remarks

This work demonstrates that fairness in federated learning is achievable without sacrificing accuracy, but requires careful algorithm selection based on data characteristics. The key insight is that fairness interventions are most effective when they match the type of bias present: coverage-based methods excel when representation bias dominates, while fairness-aware aggregation is effective when algorithmic bias is the primary concern. Local debiasing provides consistent improvements by addressing prediction bias at the client level, demonstrating strong complementarity with server-side interventions.

Our novel FedCvg-Ratio algorithm addresses a key limitation of existing coverage-based methods by eliminating manual threshold tuning while providing superior stability through EMA smoothing. The algorithm achieves substantial fairness improvements across all

heterogeneity levels, making it a robust choice for practitioners seeking to optimize multiple fairness objectives simultaneously.

As federated learning continues to gain adoption in sensitive domains such as healthcare, finance, and criminal justice, ensuring fairness becomes increasingly critical. Biased models can perpetuate and amplify existing societal inequalities, leading to discriminatory outcomes that harm vulnerable populations. Our comprehensive comparative evaluation provides practitioners with evidence-based guidelines for deploying fair federated learning systems, while our analysis of technique complementarity reveals that combining server-side and client-side interventions yields the best results. We hope this thesis serves as a practical guide for researchers and practitioners seeking to deploy equitable machine learning systems in federated settings.

Bibliography

[ABC+16] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, pages 265–283, 2016.

[Acc23] Chiara Accinelli. *Discrimination-aware data transformations*. Phd thesis, Università di Genova, 2023.

[AJJ19] A. Asudeh, Z. Jin, and H. V. Jagadish. Assessing and remedying coverage for a given dataset. In *IEEE 35th International Conference on Data Engineering (ICDE)*, pages 554–565, 2019.

[ALMK16] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica*, 2016. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[AMC20] C. Accinelli, S. Minisi, and B. Catania. Coverage-based rewriting for data preparation. In *EDBT/ICDT Workshops*, 2020.

[ASJJ21] A. Asudeh, N. Shahbazi, Z. Jin, and H. V. Jagadish. Identifying insufficient data coverage for ordinal continuous-valued attributes. In *Proceedings of the International Conference on Management of Data (SIGMOD)*, pages 129–141, 2021.

[BEG+19] Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloe Kiddon, Jakub Konečnỳ, Stefano Mazzocchi, H Brendan McMahan, et al. Towards federated learning at scale: System design. In *Proceedings of Machine Learning and Systems*, volume 1, pages 374–388, 2019.

[BFH+18] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas,

Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL: https://github.com/google/jax.

[Bro23] Martina Brocchi. Representation bias mitigation in federated learning. Master's thesis, Universita degli studi di Genova, 2023.

[BS16] Solon Barocas and Andrew D Selbst. Big data's disparate impact. *California Law Review*, 104:671–732, 2016. URL: https://www.californialawreview.org/print/big-datas-disparate-impact/.

[BTM+20] Daniel J Beutel, Taner Topal, Akhil Mathur, Xinchi Qiu, Javier Fernandez-Marques, Yan Gao, Lorenzo Sani, Kwing Hei Li, Titouan Parcollet, Pedro Porto Buarque de Gusmão, and Nicholas D Lane. Flower: A friendly federated learning research framework. *arXiv preprint arXiv:2007.14390*, 2020. URL: https://arxiv.org/abs/2007.14390.

[CBHK02] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002. doi:10.1613/jair.953.

[CWJ22] Yae Jee Cho, Jianyu Wang, and Gauri Joshi. Towards understanding biased client selection in federated learning. In *International Conference on Artificial Intelligence and Statistics*, pages 10351–10375. PMLR, 2022.

[CWV+17] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, volume 30, pages 3992–4001, 2017. URL: https://proceedings.neurips.cc/paper/2017/hash/9a49a25d845a483fae4be7e341368e36-Abstract.html.

[DHP+12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226, 2012. URL: https://dl.acm.org/doi/10.1145/2090236.2090255, doi:10.1145/2090236.2090255.

[EYH+23] Yahya H Ezzeldin, Shen Yan, Chaoyang He, Emilio Ferrara, and A Salman Avestimehr. Fairfed: Enabling group fairness in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 7494–7502, 2023. URL: https://ojs.aaai.org/index.php/AAAI/article/view/25911, doi:10.1609/aaai.v37i6.25911.

[HLS+20] Chaoyang He, Songze Li, Jinhyun So, Xiao Zeng, Mi Zhang, Hongyi Wang, Xiaoyang Wang, Praneeth Vepakomma, Abhishek Singh, Hang Qiu, et al.

Fedml: A research library and benchmark for federated machine learning. *arXiv preprint arXiv:2007.13518*, 2020. URL: https://arxiv.org/abs/2007.13518.

[HPS16] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, volume 29, pages 3323–3331, 2016. Also: arXiv:1610.02413. URL: https://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning.

[HQB19] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019. URL: https://arxiv.org/abs/1909.06335.

[HRM+18] Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*, 2018. URL: https://arxiv.org/abs/1811.03604.

[Hun86] J. S. Hunter. The exponentially weighted moving average. *Journal of Quality Technology*, 18(4):203–210, 1986. doi:10.1080/00224065.1986.11979014.

[KB96] Ron Kohavi and Barry Becker. Adult data set. UCI Machine Learning Repository, 1996. URL: https://archive.ics.uci.edu/ml/datasets/adult.

[KC12] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012. doi:10.1007/s10115-011-0463-8.

[KMA+21] Peter Kairouz, H Brendan McMahan, Brendan Avent, et al. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2):1–210, 2021. doi:10.1561/2200000083.

[LGAJ20] Y. Lin, Y. Guan, A. Asudeh, and H. V. Jagadish. Identifying insufficient data coverage in databases with multiple relations. *Proceedings of the VLDB Endowment*, 13(12):2229–2242, 2020.

[LSBS19] Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. *arXiv preprint arXiv:1905.10497*, 2019. URL: https://arxiv.org/abs/1905.10497.

[LSZ+20] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In

Proceedings of Machine Learning and Systems (MLSys), volume 2, pages 429–450, 2020. URL: https://proceedings.mlsys.org/paper/2020/hash/1f5fe83998a09396ebe6477d9475ba0c-Abstract.html.

[LXWY20] Lingjuan Lyu, Xinyi Xu, Qian Wang, and Han Yu. Collaborative fairness in federated learning. In *Federated Learning: Privacy and Incentive*, pages 189–204. Springer, 2020. doi:10.1007/978-3-030-63076-8_14.

[MMR+17] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, 2017. URL: https://proceedings.mlr.press/v54/mcmahan17a.html.

[MMS+21] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6):1–35, 2021. doi:10.1145/3457607.

[MPKF20] L. Mazilu, N. W. Paton, N. Konstantinou, and A. A. A. Fernandes. Fairness in data wrangling. In *International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 341–348, 2020.

[MSS19] Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *International Conference on Machine Learning*, pages 4615–4625. PMLR, 2019. URL: https://proceedings.mlr.press/v97/mohri19a.html.

[NAJ22] F. Nargesian, A. Asudeh, and H. V. Jagadish. Responsible data integration: Next-generation challenges. In *Proceedings of the International Conference on Management of Data (SIGMOD)*, pages 2458–2464, 2022.

[OCDK19] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2:13, 2019. doi:10.3389/fdata.2019.00013.

[Ope22] OpenML. Acsincome. https://www.openml.org/d/43136, 2022. OpenML dataset ID: 43136.

[PGM+19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[Pro16] ProPublica. COMPAS recidivism risk score data and analysis. ProPublica, 2016. URL: https://github.com/propublica/compas-analysis.

[RHL+20] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, et al. The future of digital health with federated learning. *NPJ Digital Medicine*, 3(1):119, 2020. doi:10.1038/s41746-020-00323-1.

[RLWS21] Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh. Fairbatch: Batch selection for model fairness. In *International Conference on Learning Representations*, 2021. URL: https://openreview.net/forum?id=YNnpaAKeCfx.

[RTD+18] Theo Ryffel, Andrew Trask, Morten Dahl, Bobby Wagner, Jason Mancuso, Daniel Rueckert, and Jonathan Passerat-Palmbach. A generic framework for privacy preserving deep learning. *arXiv preprint arXiv:1811.04017*, 2018. URL: https://arxiv.org/abs/1811.04017.

[SACA26] Teresa Salazar, Helder Araujo, Alberto Cano, and Pedro Henriques Abreu. A survey on group fairness in federated learning: challenges, taxonomy of solutions and directions for future research. *Artificial Intelligence Review*, 59(81), 2026. doi:10.1007/s10462-025-11475-5.

[SARS23] Nima Shahbazi, Ehsan Abbasnejad, Damith C. Ranasinghe, and Mahsa Salehi. Representation bias in data: A survey on identification and resolution techniques. *ACM Computing Surveys*, 55(13s):293:1–293:39, July 2023. doi:10.1145/3588433.

[SFAA23] Teresa Salazar, Miguel Fernandes, Helder Araújo, and Pedro Henriques Abreu. Fair-fate: Fair federated learning with momentum. *arXiv preprint arXiv:2310.10049*, 2023. URL: https://arxiv.org/abs/2310.10049.

[SG19] Harini Suresh and John V Guttag. A framework for understanding sources of harm throughout the machine learning life cycle. *arXiv preprint arXiv:1901.10002*, 2019. URL: https://arxiv.org/abs/1901.10002.

[SHS19] B. Salimi, B. Howe, and D. Suciu. Data management for causal algorithmic fairness. *IEEE Data Engineering Bulletin*, 42(3):24–35, 2019.

[SKTH25] Takumi Suimon, Yuki Koizumi, Junji Takemasa, and Toru Hasegawa. Per-element secure aggregation against data reconstruction attacks in federated learning. *arXiv preprint arXiv:2508.04285*, 2025. Version 2.

[SLAJ23] Nima Shahbazi, Yin Lin, Abolfazl Asudeh, and H. V. Jagadish. Representation bias in data: A survey on identification and resolution techniques. *ACM Comput. Surv.*, 55(13s):293:1–293:39, 2023. URL: https://arxiv.org/abs/2203.11852.

[SSAD22] S. Shetiya, I. P. Swift, A. Asudeh, and G. Das. Fairness-aware range queries for selecting unbiased data. In *IEEE 38th International Conference on Data Engineering (ICDE)*, pages 1423–1436, 2022.

[TRO+19] K. H. Tae, Y. Roh, Y. H. Oh, H. Kim, and S. E. Whang. Data cleaning for accurate, fair, and robust models: A big data-ai integration approach. In *International Workshop on Data Management for End-to-End Machine Learning*, pages 1–4, 2019.

[TW21] K. H. Tae and S. E. Whang. Slice tuner: A selective data acquisition framework for accurate and fair machine learning models, 2021.

[VLA19] I. Valentim, N. Lourenço, and N. Antunes. The impact of data preparation on the fairness of software systems. In *IEEE International Symposium on Software Reliability Engineering (ISSRE)*, pages 391–401, 2019.

[VR18] Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness (FairWare)*, pages 1–7, 2018. doi:10.1145/3194770.3194776.

[WKNL20] Hao Wang, Zachary Kaplan, Di Niu, and Baochun Li. Optimizing federated learning on non-iid data with reinforcement learning. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*, pages 1698–1707. IEEE, 2020. Proposes FAVOR, a client selection framework using deep reinforcement learning. doi:10.1109/INFOCOM41043.2020.9155494.

[YH21] A. Yan and B. Howe. Equitensors: Learning fair integrations of heterogeneous urban data. In *Proceedings of the International Conference on Management of Data (SIGMOD)*, pages 2338–2347, 2021.

[YWZ+20] Miao Yang, Akitanoshou Wong, Hongbin Zhu, Haifeng Wang, and Hua Qian. Federated learning with class imbalance reduction. *arXiv preprint arXiv:2011.11266*, 2020. URL: https://arxiv.org/abs/2011.11266.

[ZCL22] Yuchen Zeng, Hongxu Chen, and Kangwook Lee. Improving fairness via federated learning. *arXiv preprint arXiv:2110.15545*, 2022. URL: https://arxiv.org/abs/2110.15545.

[ZKW20] D. Zhang, Z. Kou, and D. Wang. Fairfl: A fair federated learning approach to reducing demographic bias in privacy-sensitive classification models. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 1051–1060. IEEE, 2020. doi:10.1109/BigData50022.2020.9378043.

[ZLM18] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018. doi:10.1145/3278721.3278779.

[ZWS+13] Rich Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333. PMLR, 2013. URL: https://proceedings.mlr.press/v28/zemel13.html.

Appendix A

Dirichlet Partitioning Results: Adult Dataset

Table A.1: Adult dataset: Performance at $\alpha = 0.1$ (best values across learning rates, averaged over 5 seeds)

Method	Acc	EOD	SPD	AOD	AccDiff	Prec
FedAvg	0.838	0.104	0.151	0.083	0.122	0.748
FedAvg+LD	0.846	0.065	0.142	0.044	0.115	0.735
FedAvg+PS(0.5)	0.830	0.100	0.135	0.076	0.126	0.748
FedAvg+PS(0.7)	0.831	0.105	0.136	0.079	0.125	0.747
FedAvg+LD+PS(0.5)	0.846	0.065	0.140	0.042	0.116	0.739
FedAvg+LD+PS(0.7)	**0.846**	0.062	0.140	0.042	0.115	0.737
FairFed(EOD)	0.815	0.089	0.104	0.064	0.139	0.740
FairFed(SPD)	0.818	0.103	0.117	0.075	0.134	0.712
FairFed(AccDiff)	0.816	0.090	0.108	0.065	0.137	0.732
FairFed(EOD)+LD	0.831	0.039	0.117	0.033	0.131	0.732
FairFed(SPD)+LD	0.823	0.052	0.107	0.038	0.131	0.720
FairFed(AccDiff)+LD	0.824	0.042	0.103	0.032	0.130	0.727
FairFed(EOD)+PS(0.5)	0.831	0.094	0.132	0.073	0.130	0.759
FairFed(SPD)+PS(0.5)	0.825	0.095	0.124	0.072	0.135	0.769
FairFed(AccDiff)+PS(0.5)	0.820	0.100	0.113	0.071	0.133	0.756
FairFed(EOD)+PS(0.7)	0.825	0.093	0.123	0.071	0.135	0.763
FairFed(SPD)+PS(0.7)	0.818	0.094	0.108	0.067	0.137	0.759
FairFed(AccDiff)+PS(0.7)	0.829	0.094	0.130	0.072	0.128	0.754
FairFed(EOD)+LD+PS(0.5)	0.831	0.032	0.111	0.027	0.130	0.759
FairFed(SPD)+LD+PS(0.5)	0.822	0.057	0.100	0.038	0.133	0.773
FairFed(AccDiff)+LD+PS(0.5)	0.834	0.036	0.113	0.026	0.125	0.751
FairFed(EOD)+LD+PS(0.7)	0.823	0.031	0.098	0.031	0.135	0.770

Continued on next page

Table A.1 – continued from previous page

Method	Acc	EOD	SPD	AOD	AccDiff	Prec
FairFed(SPD)+LD+PS(0.7)	0.823	0.050	0.098	0.036	0.131	0.754
FairFed(AccDiff)+LD+PS(0.7)	0.839	0.043	0.132	0.042	0.118	0.742
FedCvg(0.0001)	0.838	0.097	0.147	0.077	0.120	0.754
FedCvg(0.01)	0.793	0.030	0.053	0.024	0.164	0.837
FedCvg(0.0001)+LD	0.845	0.051	0.136	0.039	**0.114**	0.739
FedCvg(0.01)+LD	0.791	**0.015**	0.046	0.015	0.166	0.836
FedCvg(0.0001)+PS(0.5)	0.839	0.104	0.150	0.082	0.118	0.751
FedCvg(0.01)+PS(0.5)	0.791	0.053	0.051	0.035	0.162	0.828
FedCvg(0.0001)+PS(0.7)	0.838	0.101	0.148	0.080	0.120	0.753
FedCvg(0.01)+PS(0.7)	0.788	0.044	0.047	0.030	0.166	0.840
FedCvg(0.0001)+LD+PS(0.5)	0.845	0.047	0.140	0.039	**0.114**	0.739
FedCvg(0.01)+LD+PS(0.5)	0.830	0.029	0.109	0.033	0.130	0.768
FedCvg(0.0001)+LD+PS(0.7)	0.844	0.051	0.136	0.037	0.115	0.738
FedCvg(0.01)+LD+PS(0.7)	0.828	0.018	0.102	0.026	0.132	0.778
FedCvg-Ratio(0.5)	0.782	0.031	0.030	0.018	0.170	0.861
FedCvg-Ratio(0.9)	0.782	0.031	0.030	0.018	0.170	0.861
FedCvg-Ratio(0.5)+LD	0.782	0.025	**0.027**	**0.013**	0.172	0.863
FedCvg-Ratio(0.9)+LD	0.782	0.025	**0.027**	**0.013**	0.172	0.863
FedCvg-Ratio(0.5)+PS(0.5)	0.786	0.038	0.038	0.023	0.166	0.854
FedCvg-Ratio(0.9)+PS(0.5)	0.785	0.036	0.036	0.022	0.167	0.854
FedCvg-Ratio(0.5)+PS(0.7)	0.785	0.037	0.036	0.023	0.168	0.851
FedCvg-Ratio(0.9)+PS(0.7)	0.785	0.035	0.036	0.022	0.168	0.852
FedCvg-Ratio(0.5)+LD+PS(0.5)	0.784	0.019	0.030	**0.013**	0.170	0.862
FedCvg-Ratio(0.9)+LD+PS(0.5)	0.795	0.027	0.044	0.018	0.160	**0.871**
FedCvg-Ratio(0.5)+LD+PS(0.7)	0.784	0.026	0.032	0.016	0.169	0.851
FedCvg-Ratio(0.9)+LD+PS(0.7)	0.785	0.026	0.033	0.016	0.169	0.868

Table A.2: Adult dataset: Performance at $\alpha = 0.2$ (best values across learning rates, averaged over 5 seeds)

Method	Acc	EOD	SPD	AOD	AccDiff	Prec
FedAvg	**0.850**	0.087	0.171	0.080	**0.118**	0.739
FedAvg+LD	0.846	0.076	0.133	0.040	0.118	0.733
FedAvg+PS(0.5)	**0.850**	0.082	0.170	0.077	**0.118**	0.739
FedAvg+PS(0.7)	**0.850**	0.082	0.170	0.077	**0.118**	0.739
FedAvg+LD+PS(0.5)	0.830	0.072	0.108	0.038	0.130	0.752
FedAvg+LD+PS(0.7)	0.831	0.077	0.107	0.042	0.128	0.754
FairFed(EOD)	0.815	0.088	0.100	0.062	0.143	0.785
FairFed(SPD)	0.827	0.094	0.125	0.071	0.133	0.773
FairFed(AccDiff)	0.817	0.083	0.103	0.060	0.140	0.774

Continued on next page

Table A.2 – continued from previous page

Method	Acc	EOD	SPD	AOD	AccDiff	Prec
FairFed(EOD)+LD	0.814	0.072	0.089	0.042	0.141	0.762
FairFed(SPD)+LD	0.819	0.069	0.091	0.043	0.140	0.773
FairFed(AccDiff)+LD	0.816	0.064	0.097	0.050	0.143	0.773
FairFed(EOD)+PS(0.5)	0.818	0.088	0.105	0.063	0.139	0.771
FairFed(SPD)+PS(0.5)	0.827	0.088	0.123	0.067	0.133	0.774
FairFed(AccDiff)+PS(0.5)	0.829	0.095	0.127	0.072	0.131	0.770
FairFed(EOD)+PS(0.7)	0.827	0.092	0.124	0.070	0.133	0.771
FairFed(SPD)+PS(0.7)	0.839	0.091	0.149	0.076	0.126	0.759
FairFed(AccDiff)+PS(0.7)	0.826	0.086	0.125	0.067	0.133	0.751
FairFed(EOD)+LD+PS(0.5)	0.816	0.073	0.100	0.055	0.142	0.770
FairFed(SPD)+LD+PS(0.5)	0.809	0.069	0.083	0.048	0.148	0.785
FairFed(AccDiff)+LD+PS(0.5)	0.829	0.065	0.104	0.036	0.133	0.759
FairFed(EOD)+LD+PS(0.7)	0.826	0.060	0.109	0.041	0.135	0.762
FairFed(SPD)+LD+PS(0.7)	0.827	0.067	0.113	0.045	0.133	0.755
FairFed(AccDiff)+LD+PS(0.7)	0.826	0.068	0.111	0.046	0.134	0.758
FedCvg(0.0001)	0.829	0.097	0.129	0.073	0.126	0.766
FedCvg(0.01)	0.772	0.016	**0.017**	**0.009**	0.181	0.859
FedCvg(0.0001)+LD	0.833	0.053	0.112	0.029	0.123	0.763
FedCvg(0.01)+LD	0.796	**0.013**	0.049	0.011	0.162	0.824
FedCvg(0.0001)+PS(0.5)	0.828	0.091	0.125	0.069	0.128	0.770
FedCvg(0.01)+PS(0.5)	0.775	0.044	0.026	0.025	0.175	0.860
FedCvg(0.0001)+PS(0.7)	0.828	0.088	0.124	0.067	0.129	0.778
FedCvg(0.01)+PS(0.7)	0.770	0.031	0.018	0.017	0.180	**0.861**
FedCvg(0.0001)+LD+PS(0.5)	0.832	0.051	0.111	0.026	0.124	0.769
FedCvg(0.01)+LD+PS(0.5)	0.800	0.020	0.060	0.018	0.156	0.836
FedCvg(0.0001)+LD+PS(0.7)	0.832	0.041	0.111	0.025	0.124	0.770
FedCvg(0.01)+LD+PS(0.7)	0.813	**0.013**	0.074	0.018	0.148	0.821
FedCvg-Ratio(0.5)	0.813	0.036	0.079	0.028	0.145	0.835
FedCvg-Ratio(0.9)	0.813	0.036	0.079	0.028	0.145	0.835
FedCvg-Ratio(0.5)+LD	0.812	0.031	0.071	0.016	0.146	0.838
FedCvg-Ratio(0.9)+LD	0.812	0.031	0.071	0.016	0.146	0.838
FedCvg-Ratio(0.5)+PS(0.5)	0.818	0.042	0.086	0.023	0.141	0.810
FedCvg-Ratio(0.9)+PS(0.5)	0.817	0.036	0.088	0.025	0.142	0.810
FedCvg-Ratio(0.5)+PS(0.7)	0.820	0.041	0.093	0.028	0.139	0.801
FedCvg-Ratio(0.9)+PS(0.7)	0.820	0.042	0.094	0.028	0.139	0.800
FedCvg-Ratio(0.5)+LD+PS(0.5)	0.814	0.025	0.077	0.020	0.144	0.819
FedCvg-Ratio(0.9)+LD+PS(0.5)	0.801	0.031	0.059	0.022	0.154	0.832
FedCvg-Ratio(0.5)+LD+PS(0.7)	0.815	0.024	0.083	0.024	0.141	0.803
FedCvg-Ratio(0.9)+LD+PS(0.7)	0.804	0.036	0.066	0.026	0.152	0.827

Table A.3: Adult dataset: Performance at $\alpha = 0.5$ (best values across learning rates, averaged over 5 seeds)

Method	Acc	EOD	SPD	AOD	AccDiff	Prec
FedAvg	0.829	0.087	0.131	0.070	0.129	0.750
FedAvg+LD	0.813	0.069	0.094	0.051	0.140	0.760
FedAvg+PS(0.5)	0.824	0.067	0.119	0.058	0.137	0.765
FedAvg+PS(0.7)	0.824	0.067	0.119	0.058	0.137	0.766
FedAvg+LD+PS(0.5)	0.799	0.054	0.065	0.038	0.155	0.792
FedAvg+LD+PS(0.7)	0.807	0.048	0.080	0.038	0.148	0.778
FairFed(EOD)	0.805	0.077	0.079	0.052	0.149	0.788
FairFed(SPD)	0.818	0.087	0.106	0.063	0.136	0.763
FairFed(AccDiff)	0.817	0.085	0.103	0.061	0.138	0.769
FairFed(EOD)+LD	0.790	0.059	**0.047**	0.035	0.160	0.804
FairFed(SPD)+LD	0.810	0.057	0.087	0.044	0.143	0.763
FairFed(AccDiff)+LD	0.810	0.053	0.085	0.041	0.144	0.768
FairFed(EOD)+PS(0.5)	0.813	0.067	0.095	0.052	0.146	0.787
FairFed(SPD)+PS(0.5)	0.825	0.076	0.121	0.063	0.135	0.762
FairFed(AccDiff)+PS(0.5)	0.812	0.067	0.096	0.053	0.147	0.790
FairFed(EOD)+PS(0.7)	0.813	0.072	0.097	0.055	0.145	0.783
FairFed(SPD)+PS(0.7)	0.824	0.072	0.119	0.060	0.136	0.767
FairFed(AccDiff)+PS(0.7)	0.812	0.066	0.095	0.052	0.147	0.790
FairFed(EOD)+LD+PS(0.5)	0.816	0.046	0.087	0.032	0.142	0.764
FairFed(SPD)+LD+PS(0.5)	0.808	0.043	0.080	0.036	0.148	0.780
FairFed(AccDiff)+LD+PS(0.5)	0.807	0.044	0.079	0.036	0.148	0.775
FairFed(EOD)+LD+PS(0.7)	0.816	0.043	0.087	0.032	0.143	0.762
FairFed(SPD)+LD+PS(0.7)	0.797	0.045	0.061	0.033	0.158	0.803
FairFed(AccDiff)+LD+PS(0.7)	0.807	0.040	0.078	0.034	0.149	0.780
FedCvg(0.0001)	0.808	0.113	0.087	0.071	0.141	0.775
FedCvg(0.01)	0.790	0.059	0.052	0.038	0.162	0.831
FedCvg(0.0001)+LD	**0.838**	0.064	0.117	0.040	**0.117**	0.759
FedCvg(0.01)+LD	0.797	**0.023**	0.052	0.019	0.157	**0.837**
FedCvg(0.0001)+PS(0.5)	0.804	0.106	0.079	0.066	0.146	0.790
FedCvg(0.01)+PS(0.5)	0.791	0.062	0.053	0.040	0.161	0.829
FedCvg(0.0001)+PS(0.7)	0.815	0.104	0.100	0.070	0.139	0.785
FedCvg(0.01)+PS(0.7)	0.792	0.065	0.055	0.041	0.160	0.824
FedCvg(0.0001)+LD+PS(0.5)	0.823	0.056	0.087	0.029	0.128	0.776
FedCvg(0.01)+LD+PS(0.5)	0.799	0.037	0.058	0.027	0.155	0.824
FedCvg(0.0001)+LD+PS(0.7)	0.836	0.056	0.109	0.034	0.118	0.765
FedCvg(0.01)+LD+PS(0.7)	0.798	0.030	0.055	0.023	0.156	0.829
FedCvg-Ratio(0.5)	0.808	0.044	0.073	0.034	0.148	0.819
FedCvg-Ratio(0.9)	0.808	0.044	0.073	0.034	0.148	0.819
FedCvg-Ratio(0.5)+LD	0.833	0.027	0.102	0.019	0.127	0.791
FedCvg-Ratio(0.9)+LD	0.833	0.027	0.102	0.019	0.127	0.791

Continued on next page

Table A.3 – continued from previous page

Method	Acc	EOD	SPD	AOD	AccDiff	Prec
FedCvg-Ratio(0.5)+PS(0.5)	0.822	0.045	0.097	0.040	0.137	0.802
FedCvg-Ratio(0.9)+PS(0.5)	0.823	0.046	0.097	0.040	0.136	0.803
FedCvg-Ratio(0.5)+PS(0.7)	0.809	0.042	0.073	0.033	0.148	0.819
FedCvg-Ratio(0.9)+PS(0.7)	0.808	0.041	0.073	0.033	0.149	0.818
FedCvg-Ratio(0.5)+LD+PS(0.5)	0.833	0.024	0.099	**0.017**	0.126	0.793
FedCvg-Ratio(0.9)+LD+PS(0.5)	0.831	0.026	0.100	0.024	0.129	0.791
FedCvg-Ratio(0.5)+LD+PS(0.7)	0.833	0.026	0.100	0.018	0.127	0.790
FedCvg-Ratio(0.9)+LD+PS(0.7)	0.833	0.024	0.102	0.019	0.126	0.785

Table A.4: Adult dataset: Performance at $\alpha = 10.0$ (best values across learning rates, averaged over 5 seeds)

Method	Acc	EOD	SPD	AOD	AccDiff	Prec
FedAvg	0.786	0.068	0.043	0.039	0.161	0.799
FedAvg+LD	0.795	0.038	0.045	0.019	0.156	0.799
FedAvg+PS(0.5)	0.786	0.070	0.042	0.040	0.161	0.805
FedAvg+PS(0.7)	0.785	0.070	0.041	0.039	0.162	0.807
FedAvg+LD+PS(0.5)	0.795	0.037	0.042	0.018	0.157	0.801
FedAvg+LD+PS(0.7)	0.794	0.038	0.041	0.019	0.157	0.801
FairFed(EOD)	0.786	0.070	0.043	0.040	0.161	0.805
FairFed(SPD)	0.787	0.073	0.044	0.042	0.160	0.805
FairFed(AccDiff)	0.801	0.075	0.073	0.049	0.149	0.781
FairFed(EOD)+LD	0.795	0.040	0.043	0.020	0.156	0.799
FairFed(SPD)+LD	0.795	0.041	0.043	0.020	0.156	0.800
FairFed(AccDiff)+LD	0.798	0.044	0.048	0.022	0.153	0.802
FairFed(EOD)+PS(0.5)	0.787	0.076	0.045	0.043	0.160	0.803
FairFed(SPD)+PS(0.5)	0.786	0.071	0.042	0.040	0.161	0.805
FairFed(AccDiff)+PS(0.5)	0.786	0.068	0.042	0.039	0.162	0.804
FairFed(EOD)+PS(0.7)	0.787	0.075	0.045	0.043	0.160	0.802
FairFed(SPD)+PS(0.7)	0.786	0.070	0.042	0.040	0.162	0.806
FairFed(AccDiff)+PS(0.7)	0.786	0.071	0.042	0.040	0.162	0.808
FairFed(EOD)+LD+PS(0.5)	0.795	0.042	0.043	0.021	0.156	0.799
FairFed(SPD)+LD+PS(0.5)	0.794	0.039	0.042	0.019	0.157	0.801
FairFed(AccDiff)+LD+PS(0.5)	0.794	0.040	0.042	0.019	0.157	0.802
FairFed(EOD)+LD+PS(0.7)	0.795	0.045	0.042	0.022	0.156	0.800
FairFed(SPD)+LD+PS(0.7)	0.794	0.038	0.041	0.019	0.157	0.801
FairFed(AccDiff)+LD+PS(0.7)	0.794	0.039	0.041	0.019	0.157	0.801
FedCvg(0.0001)	0.789	0.070	0.044	0.040	0.162	0.818
FedCvg(0.01)	0.784	0.057	0.035	0.032	0.169	**0.836**
FedCvg(0.0001)+LD	0.786	0.035	0.033	0.020	0.168	0.831
FedCvg(0.01)+LD	0.790	0.024	0.041	0.018	0.167	0.823

Continued on next page

Table A.4 – continued from previous page

Method	Acc	EOD	SPD	AOD	AccDiff	Prec
FedCvg(0.0001)+PS(0.5)	0.802	0.065	0.069	0.044	0.154	0.804
FedCvg(0.01)+PS(0.5)	0.784	0.057	0.035	0.032	0.169	0.834
FedCvg(0.0001)+PS(0.7)	0.787	0.063	0.041	0.036	0.165	0.818
FedCvg(0.01)+PS(0.7)	0.784	0.057	0.035	0.033	0.169	0.833
FedCvg(0.0001)+LD+PS(0.5)	0.786	0.033	0.032	0.019	0.169	0.834
FedCvg(0.01)+LD+PS(0.5)	0.791	0.023	0.042	0.018	0.166	0.823
FedCvg(0.0001)+LD+PS(0.7)	0.784	0.031	**0.031**	0.019	0.170	0.831
FedCvg(0.01)+LD+PS(0.7)	0.791	0.022	0.041	0.017	0.167	0.824
FedCvg-Ratio(0.5)	**0.836**	0.032	0.120	0.029	**0.127**	0.758
FedCvg-Ratio(0.9)	**0.836**	0.032	0.120	0.029	**0.127**	0.758
FedCvg-Ratio(0.5)+LD	0.814	0.021	0.078	0.019	0.143	0.774
FedCvg-Ratio(0.9)+LD	0.814	0.021	0.078	0.019	0.143	0.774
FedCvg-Ratio(0.5)+PS(0.5)	0.836	0.031	0.120	0.029	0.128	0.759
FedCvg-Ratio(0.9)+PS(0.5)	**0.836**	0.035	0.119	0.027	0.128	0.759
FedCvg-Ratio(0.5)+PS(0.7)	**0.836**	0.031	0.120	0.029	0.128	0.760
FedCvg-Ratio(0.9)+PS(0.7)	0.836	0.032	0.119	0.028	0.128	0.759
FedCvg-Ratio(0.5)+LD+PS(0.5)	0.804	0.020	0.059	**0.014**	0.153	0.793
FedCvg-Ratio(0.9)+LD+PS(0.5)	0.814	0.019	0.077	0.017	0.144	0.776
FedCvg-Ratio(0.5)+LD+PS(0.7)	0.827	0.015	0.099	0.022	0.132	0.761
FedCvg-Ratio(0.9)+LD+PS(0.7)	0.827	**0.014**	0.098	0.021	0.132	0.762

Table A.5: Adult dataset: Performance at $\alpha = 5000.0$ (best values across learning rates, averaged over 5 seeds)

Method	Acc	EOD	SPD	AOD	AccDiff	Prec
FedAvg	0.785	0.068	0.040	0.038	0.163	0.794
FedAvg+LD	0.782	0.039	**0.030**	0.021	0.168	0.804
FedAvg+PS(0.5)	0.784	0.067	0.040	0.038	0.164	0.796
FedAvg+PS(0.7)	0.784	0.066	0.040	0.037	0.164	0.796
FedAvg+LD+PS(0.5)	0.782	0.037	**0.029**	0.021	0.168	**0.804**
FedAvg+LD+PS(0.7)	0.782	0.038	**0.029**	0.021	0.168	**0.804**
FairFed(EOD)	0.784	0.069	0.040	0.039	0.163	0.794
FairFed(SPD)	0.785	0.069	0.041	0.039	0.163	0.795
FairFed(AccDiff)	0.785	0.067	0.041	0.038	0.163	0.794
FairFed(EOD)+LD	0.781	0.039	**0.029**	0.022	0.168	0.803
FairFed(SPD)+LD	0.782	0.037	**0.029**	0.021	0.168	0.802
FairFed(AccDiff)+LD	0.782	0.038	0.030	0.021	0.168	**0.804**
FairFed(EOD)+PS(0.5)	0.785	0.068	0.040	0.038	0.163	0.795
FairFed(SPD)+PS(0.5)	0.784	0.068	0.040	0.038	0.164	0.796
FairFed(AccDiff)+PS(0.5)	0.785	0.068	0.040	0.038	0.163	0.797
FairFed(EOD)+PS(0.7)	0.784	0.067	0.040	0.038	0.164	0.794

Continued on next page

Table A.5 – continued from previous page

Method	Acc	EOD	SPD	AOD	AccDiff	Prec
FairFed(SPD)+PS(0.7)	0.784	0.067	0.040	0.038	0.164	0.796
FairFed(AccDiff)+PS(0.7)	0.784	0.067	0.040	0.038	0.164	0.797
FairFed(EOD)+LD+PS(0.5)	0.782	0.039	**0.030**	0.022	0.168	0.803
FairFed(SPD)+LD+PS(0.5)	0.782	0.037	**0.029**	0.021	0.168	0.803
FairFed(AccDiff)+LD+PS(0.5)	0.782	0.037	**0.029**	0.021	0.168	**0.804**
FairFed(EOD)+LD+PS(0.7)	0.782	0.038	**0.030**	0.021	0.168	0.803
FairFed(SPD)+LD+PS(0.7)	0.782	0.038	**0.029**	0.021	0.168	**0.804**
FairFed(AccDiff)+LD+PS(0.7)	0.782	0.038	**0.029**	0.021	0.168	**0.804**
FedCvg(0.0001)	0.782	0.069	0.042	0.039	0.165	0.794
FedCvg(0.01)	0.782	0.069	0.041	0.039	0.165	0.794
FedCvg(0.0001)+LD	0.779	0.043	0.031	0.024	0.170	0.800
FedCvg(0.01)+LD	0.779	0.043	0.031	0.024	0.170	0.802
FedCvg(0.0001)+PS(0.5)	0.782	0.069	0.041	0.039	0.165	0.794
FedCvg(0.01)+PS(0.5)	0.781	0.068	0.041	0.038	0.165	0.796
FedCvg(0.0001)+PS(0.7)	0.782	0.068	0.041	0.039	0.165	0.798
FedCvg(0.01)+PS(0.7)	0.782	0.068	0.041	0.039	0.165	0.797
FedCvg(0.0001)+LD+PS(0.5)	0.779	0.043	0.031	0.023	0.170	0.802
FedCvg(0.01)+LD+PS(0.5)	0.779	0.044	0.031	0.024	0.170	0.802
FedCvg(0.0001)+LD+PS(0.7)	0.779	0.043	0.031	0.023	0.170	0.802
FedCvg(0.01)+LD+PS(0.7)	0.779	0.044	0.031	0.024	0.170	0.804
FedCvg-Ratio(0.5)	0.821	0.042	0.104	0.040	0.135	0.759
FedCvg-Ratio(0.9)	0.821	0.042	0.104	0.040	0.135	0.759
FedCvg-Ratio(0.5)+LD	0.832	**0.012**	0.107	0.016	0.124	0.755
FedCvg-Ratio(0.9)+LD	0.832	**0.012**	0.107	0.016	0.124	0.755
FedCvg-Ratio(0.5)+PS(0.5)	0.821	0.042	0.102	0.039	0.135	0.762
FedCvg-Ratio(0.9)+PS(0.5)	0.821	0.040	0.103	0.038	0.135	0.761
FedCvg-Ratio(0.5)+PS(0.7)	0.834	0.040	0.124	0.040	0.123	0.751
FedCvg-Ratio(0.9)+PS(0.7)	0.821	0.041	0.102	0.038	0.135	0.763
FedCvg-Ratio(0.5)+LD+PS(0.5)	**0.842**	0.016	0.124	0.019	**0.115**	0.745
FedCvg-Ratio(0.9)+LD+PS(0.5)	0.832	0.013	0.105	**0.014**	0.125	0.758
FedCvg-Ratio(0.5)+LD+PS(0.7)	0.832	0.017	0.105	**0.014**	0.125	0.757
FedCvg-Ratio(0.9)+LD+PS(0.7)	0.832	0.016	0.105	**0.014**	0.125	0.758

Appendix B

Dirichlet Partitioning Results: COMPAS Dataset

Table B.1: Compas dataset: Performance at $\alpha = 0.1$ (best values across learning rates, averaged over 5 seeds)

Method	Acc	EOD	SPD	AOD	AccDiff	Prec
FedAvg	0.535	0.059	0.053	0.043	-0.024	0.520
FedAvg+LD	0.535	0.059	0.053	0.043	-0.024	0.520
FedAvg+PS(0.5)	0.533	0.062	0.056	0.046	-0.023	0.518
FedAvg+PS(0.7)	0.533	0.056	0.053	0.042	-0.021	0.518
FedAvg+LD+PS(0.5)	0.534	0.060	0.056	0.045	-0.024	0.518
FedAvg+LD+PS(0.7)	0.533	0.063	0.056	0.046	-0.025	0.516
FairFed(EOD)	0.553	0.063	0.069	0.053	-0.025	0.548
FairFed(SPD)	0.533	0.057	0.051	0.041	-0.020	0.510
FairFed(AccDiff)	0.531	0.055	0.051	0.041	-0.023	0.507
FairFed(EOD)+LD	0.531	0.064	0.055	0.046	-0.022	0.511
FairFed(SPD)+LD	0.533	0.056	0.053	0.042	-0.023	0.509
FairFed(AccDiff)+LD	0.531	0.055	0.050	0.040	-0.022	0.507
FairFed(EOD)+PS(0.5)	0.551	0.058	0.066	0.049	-0.026	0.546
FairFed(SPD)+PS(0.5)	0.533	0.061	0.051	0.041	-0.028	0.511
FairFed(AccDiff)+PS(0.5)	0.531	0.059	0.049	0.040	-0.026	0.507
FairFed(EOD)+PS(0.7)	0.533	0.067	0.060	0.050	-0.023	0.517
FairFed(SPD)+PS(0.7)	0.535	0.058	0.045	0.037	-0.039	0.511
FairFed(AccDiff)+PS(0.7)	0.534	0.060	0.044	0.034	-0.028	0.504
FairFed(EOD)+LD+PS(0.5)	0.532	0.063	0.056	0.046	-0.023	0.515
FairFed(SPD)+LD+PS(0.5)	0.531	0.058	0.043	0.034	-0.032	0.508
FairFed(AccDiff)+LD+PS(0.5)	0.553	0.058	0.064	0.046	-0.022	0.536

Continued on next page

Table B.1 – continued from previous page

Method	Acc	EOD	SPD	AOD	AccDiff	Prec
FairFed(EOD)+LD+PS(0.7)	0.535	0.061	0.057	0.047	-0.025	0.515
FairFed(SPD)+LD+PS(0.7)	0.535	0.065	0.047	0.039	-0.032	0.511
FairFed(AccDiff)+LD+PS(0.7)	0.533	0.061	0.048	0.038	-0.029	0.506
FedCvg(0.0001)	0.536	0.034	0.031	0.019	-0.040	0.542
FedCvg(0.01)	0.537	0.039	0.031	0.020	-0.034	0.540
FedCvg(0.0001)+LD	0.535	0.036	0.025	0.020	-0.034	0.542
FedCvg(0.01)+LD	0.537	0.039	0.031	0.020	-0.034	0.540
FedCvg(0.0001)+PS(0.5)	0.534	0.036	0.030	0.023	-0.042	0.537
FedCvg(0.01)+PS(0.5)	0.536	0.033	0.030	0.020	-0.039	0.540
FedCvg(0.0001)+PS(0.7)	0.535	0.039	0.030	0.024	-0.040	0.539
FedCvg(0.01)+PS(0.7)	0.535	0.036	0.023	0.021	-0.031	0.541
FedCvg(0.0001)+LD+PS(0.5)	0.534	0.035	0.029	0.024	-0.041	0.536
FedCvg(0.01)+LD+PS(0.5)	0.536	0.035	0.030	0.020	-0.036	0.541
FedCvg(0.0001)+LD+PS(0.7)	0.536	0.037	0.037	0.025	-0.045	0.538
FedCvg(0.01)+LD+PS(0.7)	0.536	0.036	0.031	0.019	-0.036	0.540
FedCvg-Ratio(0.5)	0.554	0.029	0.025	0.018	-0.059	0.612
FedCvg-Ratio(0.9)	0.554	0.029	0.025	0.018	-0.059	0.612
FedCvg-Ratio(0.5)+LD	0.554	0.028	0.025	0.018	**-0.060**	0.612
FedCvg-Ratio(0.9)+LD	0.554	0.028	0.025	0.018	**-0.060**	0.612
FedCvg-Ratio(0.5)+PS(0.5)	0.554	0.028	0.022	0.017	-0.057	0.609
FedCvg-Ratio(0.9)+PS(0.5)	0.554	**0.027**	0.022	**0.016**	-0.059	0.609
FedCvg-Ratio(0.5)+PS(0.7)	0.554	0.029	0.024	0.017	-0.057	0.610
FedCvg-Ratio(0.9)+PS(0.7)	0.554	0.028	0.023	**0.017**	-0.058	0.610
FedCvg-Ratio(0.5)+LD+PS(0.5)	**0.555**	0.029	0.023	**0.017**	-0.058	**0.613**
FedCvg-Ratio(0.9)+LD+PS(0.5)	0.543	**0.027**	**0.017**	0.017	-0.053	0.594
FedCvg-Ratio(0.5)+LD+PS(0.7)	0.554	0.029	0.023	**0.017**	-0.057	0.609
FedCvg-Ratio(0.9)+LD+PS(0.7)	0.554	0.027	0.022	0.017	-0.058	0.610

Table B.2: Compas dataset: Performance at $\alpha = 0.2$ (best values across learning rates, averaged over 5 seeds)

Method	Acc	EOD	SPD	AOD	AccDiff	Prec
FedAvg	0.512	0.091	0.060	0.058	-0.016	0.500
FedAvg+LD	0.513	0.093	0.060	0.059	-0.015	0.501
FedAvg+PS(0.5)	0.508	0.086	0.056	0.055	-0.019	0.487
FedAvg+PS(0.7)	0.510	0.088	0.055	0.054	-0.019	0.490
FedAvg+LD+PS(0.5)	0.509	0.084	0.056	0.054	-0.019	0.486
FedAvg+LD+PS(0.7)	0.509	0.086	0.057	0.056	-0.020	0.484
FairFed(EOD)	0.507	0.082	0.054	0.052	-0.019	0.469

Continued on next page

Table B.2 – continued from previous page

Method	Acc	EOD	SPD	AOD	AccDiff	Prec
FairFed(SPD)	0.506	0.082	0.057	0.056	-0.021	0.472
FairFed(AccDiff)	0.505	0.081	0.055	0.054	-0.020	0.468
FairFed(EOD)+LD	0.506	0.081	0.053	0.052	-0.020	0.468
FairFed(SPD)+LD	0.506	0.083	0.056	0.055	-0.020	0.472
FairFed(AccDiff)+LD	0.505	0.081	0.055	0.055	-0.020	0.468
FairFed(EOD)+PS(0.5)	0.508	0.084	0.054	0.053	-0.023	0.476
FairFed(SPD)+PS(0.5)	0.509	0.083	0.055	0.055	-0.024	0.475
FairFed(AccDiff)+PS(0.5)	0.506	0.078	0.052	0.052	-0.025	0.469
FairFed(EOD)+PS(0.7)	0.508	0.077	0.055	0.054	-0.027	0.472
FairFed(SPD)+PS(0.7)	0.507	0.078	0.055	0.054	-0.028	0.464
FairFed(AccDiff)+PS(0.7)	0.508	0.081	0.052	0.051	-0.023	0.473
FairFed(EOD)+LD+PS(0.5)	0.509	0.077	0.057	0.055	-0.026	0.474
FairFed(SPD)+LD+PS(0.5)	0.507	0.079	0.054	0.053	-0.024	0.471
FairFed(AccDiff)+LD+PS(0.5)	0.508	0.080	0.053	0.052	-0.025	0.475
FairFed(EOD)+LD+PS(0.7)	0.527	0.079	0.071	0.063	-0.031	0.504
FairFed(SPD)+LD+PS(0.7)	0.507	0.080	0.053	0.052	-0.026	0.469
FairFed(AccDiff)+LD+PS(0.7)	0.507	0.080	0.052	0.051	-0.026	0.473
FedCvg(0.0001)	0.523	0.065	**0.025**	0.028	-0.036	0.517
FedCvg(0.01)	0.527	0.057	0.028	0.030	-0.034	0.522
FedCvg(0.0001)+LD	0.523	0.065	0.025	0.029	-0.036	0.516
FedCvg(0.01)+LD	0.526	0.058	0.027	0.029	-0.034	0.522
FedCvg(0.0001)+PS(0.5)	0.520	0.068	0.025	0.030	-0.037	0.512
FedCvg(0.01)+PS(0.5)	0.523	0.064	0.026	0.031	-0.037	0.518
FedCvg(0.0001)+PS(0.7)	0.520	0.070	0.026	0.032	-0.039	0.510
FedCvg(0.01)+PS(0.7)	0.523	0.066	**0.024**	0.030	-0.036	0.523
FedCvg(0.0001)+LD+PS(0.5)	0.521	0.070	**0.024**	0.031	-0.037	0.513
FedCvg(0.01)+LD+PS(0.5)	0.522	0.070	0.026	0.031	-0.037	0.517
FedCvg(0.0001)+LD+PS(0.7)	0.519	0.069	0.025	0.031	-0.039	0.508
FedCvg(0.01)+LD+PS(0.7)	0.523	0.067	0.028	0.033	-0.037	0.521
FedCvg-Ratio(0.5)	0.553	0.054	0.041	0.036	-0.067	0.599
FedCvg-Ratio(0.9)	0.553	0.054	0.041	0.036	-0.067	0.599
FedCvg-Ratio(0.5)+LD	0.522	0.055	0.027	**0.027**	-0.052	0.517
FedCvg-Ratio(0.9)+LD	0.522	0.055	0.027	**0.027**	-0.052	0.517
FedCvg-Ratio(0.5)+PS(0.5)	0.550	0.060	0.043	0.038	-0.064	0.583
FedCvg-Ratio(0.9)+PS(0.5)	0.550	0.055	0.040	0.035	-0.065	0.581
FedCvg-Ratio(0.5)+PS(0.7)	0.558	**0.052**	0.044	0.037	-0.075	0.616
FedCvg-Ratio(0.9)+PS(0.7)	0.559	0.054	0.048	0.040	**-0.076**	0.612
FedCvg-Ratio(0.5)+LD+PS(0.5)	0.521	0.060	0.025	**0.027**	-0.047	0.504
FedCvg-Ratio(0.9)+LD+PS(0.5)	0.539	0.059	0.035	0.032	-0.048	0.549
FedCvg-Ratio(0.5)+LD+PS(0.7)	0.573	0.055	0.050	0.038	-0.070	0.628
FedCvg-Ratio(0.9)+LD+PS(0.7)	**0.577**	**0.052**	0.051	0.040	-0.075	**0.665**

Table B.3: Compas dataset: Performance at $\alpha = 0.5$ (best values across learning rates, averaged over 5 seeds)

Method	Acc	EOD	SPD	AOD	AccDiff	Prec
FedAvg	**0.524**	0.036	0.041	0.032	-0.032	0.514
FedAvg+LD	**0.524**	0.037	0.041	0.032	-0.032	0.513
FedAvg+PS(0.5)	0.522	0.036	0.043	0.033	-0.033	0.506
FedAvg+PS(0.7)	0.521	0.042	0.042	0.033	-0.033	0.505
FedAvg+LD+PS(0.5)	0.522	0.037	0.040	0.031	-0.033	0.503
FedAvg+LD+PS(0.7)	0.522	0.037	0.039	0.030	-0.033	0.507
FairFed(EOD)	0.522	0.041	0.042	0.034	-0.033	0.506
FairFed(SPD)	0.520	0.043	0.045	0.036	-0.031	0.488
FairFed(AccDiff)	0.521	0.033	0.043	0.032	-0.032	0.505
FairFed(EOD)+LD	0.521	0.041	0.043	0.034	-0.033	0.504
FairFed(SPD)+LD	0.520	0.039	0.047	0.038	-0.030	0.491
FairFed(AccDiff)+LD	0.521	0.033	0.043	0.032	-0.032	0.505
FairFed(EOD)+PS(0.5)	0.523	0.044	0.044	0.035	-0.033	0.508
FairFed(SPD)+PS(0.5)	0.521	0.045	0.044	0.035	-0.033	0.500
FairFed(AccDiff)+PS(0.5)	0.521	0.044	0.049	0.039	-0.029	0.507
FairFed(EOD)+PS(0.7)	0.521	0.047	0.042	0.033	-0.033	0.502
FairFed(SPD)+PS(0.7)	0.522	0.048	0.043	0.034	-0.032	0.502
FairFed(AccDiff)+PS(0.7)	0.522	0.043	0.047	0.038	-0.031	0.508
FairFed(EOD)+LD+PS(0.5)	0.522	0.048	0.044	0.036	-0.033	0.503
FairFed(SPD)+LD+PS(0.5)	0.522	0.047	0.044	0.036	-0.031	0.500
FairFed(AccDiff)+LD+PS(0.5)	0.521	0.040	0.046	0.036	-0.031	0.503
FairFed(EOD)+LD+PS(0.7)	0.522	0.047	0.043	0.035	-0.032	0.505
FairFed(SPD)+LD+PS(0.7)	0.521	0.047	0.044	0.035	-0.031	0.506
FairFed(AccDiff)+LD+PS(0.7)	0.521	0.045	0.048	0.039	-0.032	0.503
FedCvg(0.0001)	0.521	0.042	0.036	0.027	-0.041	0.520
FedCvg(0.01)	0.521	0.045	0.047	0.038	-0.040	0.521
FedCvg(0.0001)+LD	0.521	0.041	0.036	0.027	-0.042	0.520
FedCvg(0.01)+LD	0.520	0.045	0.047	0.038	-0.039	0.519
FedCvg(0.0001)+PS(0.5)	0.520	0.034	0.031	0.022	-0.044	0.520
FedCvg(0.01)+PS(0.5)	0.523	0.038	0.042	0.032	-0.046	0.522
FedCvg(0.0001)+PS(0.7)	0.521	0.032	0.030	0.021	-0.047	0.520
FedCvg(0.01)+PS(0.7)	0.523	0.036	0.040	0.031	-0.049	**0.525**
FedCvg(0.0001)+LD+PS(0.5)	0.520	0.033	0.031	0.021	-0.045	0.516
FedCvg(0.01)+LD+PS(0.5)	**0.524**	0.038	0.041	0.031	-0.046	0.525
FedCvg(0.0001)+LD+PS(0.7)	0.521	0.033	0.030	0.021	-0.045	0.520
FedCvg(0.01)+LD+PS(0.7)	0.523	0.035	0.039	0.030	-0.048	**0.525**
FedCvg-Ratio(0.5)	0.522	0.032	**0.027**	**0.019**	**-0.054**	0.519
FedCvg-Ratio(0.9)	0.522	0.032	**0.027**	**0.019**	**-0.054**	0.519
FedCvg-Ratio(0.5)+LD	0.522	0.033	**0.027**	**0.019**	**-0.054**	0.520

Continued on next page

Table B.3 – continued from previous page

Method	Acc	EOD	SPD	AOD	AccDiff	Prec
FedCvg-Ratio(0.9)+LD	0.522	0.033	**0.027**	**0.019**	**-0.054**	0.520
FedCvg-Ratio(0.5)+PS(0.5)	0.521	0.031	**0.028**	**0.019**	**-0.054**	0.516
FedCvg-Ratio(0.9)+PS(0.5)	0.521	0.030	0.030	0.021	-0.050	0.516
FedCvg-Ratio(0.5)+PS(0.7)	0.521	0.030	0.028	**0.019**	-0.053	0.513
FedCvg-Ratio(0.9)+PS(0.7)	0.521	0.031	0.028	0.019	-0.053	0.515
FedCvg-Ratio(0.5)+LD+PS(0.5)	0.520	0.030	0.028	0.019	-0.054	0.511
FedCvg-Ratio(0.9)+LD+PS(0.5)	0.521	**0.030**	0.028	0.020	-0.053	0.511
FedCvg-Ratio(0.5)+LD+PS(0.7)	0.520	**0.030**	0.029	0.020	-0.054	0.513
FedCvg-Ratio(0.9)+LD+PS(0.7)	0.520	0.030	0.028	0.019	**-0.054**	0.510

Table B.4: Compas dataset: Performance at $\alpha = 10.0$ (best values across learning rates, averaged over 5 seeds)

Method	Acc	EOD	SPD	AOD	AccDiff	Prec
FedAvg	0.513	0.042	0.043	0.036	-0.023	0.484
FedAvg+LD	0.513	0.042	0.043	0.036	-0.023	0.484
FedAvg+PS(0.5)	0.514	0.044	0.043	0.036	-0.022	0.485
FedAvg+PS(0.7)	0.534	0.042	0.055	0.042	-0.033	0.513
FedAvg+LD+PS(0.5)	0.514	0.045	0.043	0.036	-0.022	0.485
FedAvg+LD+PS(0.7)	0.513	0.043	**0.041**	0.034	-0.024	0.484
FairFed(EOD)	0.513	0.042	0.043	0.036	-0.023	0.484
FairFed(SPD)	0.533	0.038	0.051	0.038	-0.033	0.512
FairFed(AccDiff)	0.513	0.041	0.044	0.036	-0.022	0.483
FairFed(EOD)+LD	0.513	0.041	0.043	0.036	-0.022	0.484
FairFed(SPD)+LD	0.529	**0.036**	0.044	**0.033**	-0.029	0.505
FairFed(AccDiff)+LD	0.511	0.041	0.046	0.038	-0.023	0.482
FairFed(EOD)+PS(0.5)	0.531	0.043	0.049	0.038	-0.039	0.512
FairFed(SPD)+PS(0.5)	0.534	0.039	0.049	0.036	-0.037	0.512
FairFed(AccDiff)+PS(0.5)	0.533	0.043	0.053	0.041	-0.032	0.511
FairFed(EOD)+PS(0.7)	0.532	0.044	0.046	0.038	-0.037	0.511
FairFed(SPD)+PS(0.7)	0.535	0.041	0.050	0.041	**-0.040**	0.513
FairFed(AccDiff)+PS(0.7)	0.513	0.043	**0.041**	0.034	-0.024	0.483
FairFed(EOD)+LD+PS(0.5)	0.531	0.041	0.049	0.038	-0.034	0.510
FairFed(SPD)+LD+PS(0.5)	0.531	0.040	0.047	0.036	-0.036	0.510
FairFed(AccDiff)+LD+PS(0.5)	0.513	0.042	0.041	0.034	-0.025	0.483
FairFed(EOD)+LD+PS(0.7)	0.531	0.042	0.043	0.039	-0.039	0.509
FairFed(SPD)+LD+PS(0.7)	0.531	0.040	0.046	0.035	-0.038	0.512
FairFed(AccDiff)+LD+PS(0.7)	0.513	0.041	**0.041**	**0.033**	-0.025	0.483
FedCvg(0.0001)	0.539	0.089	0.062	0.054	-0.001	0.527
FedCvg(0.01)	0.536	0.087	0.065	0.057	-0.001	0.519

Continued on next page

Table B.4 – continued from previous page

Method	Acc	EOD	SPD	AOD	AccDiff	Prec
FedCvg(0.0001)+LD	**0.579**	0.085	0.069	0.059	-0.023	0.587
FedCvg(0.01)+LD	0.536	0.092	0.064	0.057	-0.002	0.514
FedCvg(0.0001)+PS(0.5)	0.539	0.091	0.066	0.057	-0.004	0.529
FedCvg(0.01)+PS(0.5)	0.536	0.090	0.063	0.055	-0.006	0.519
FedCvg(0.0001)+PS(0.7)	0.539	0.090	0.067	0.058	-0.005	0.529
FedCvg(0.01)+PS(0.7)	0.537	0.093	0.065	0.057	-0.005	0.519
FedCvg(0.0001)+LD+PS(0.5)	0.539	0.095	0.063	0.055	-0.004	0.526
FedCvg(0.01)+LD+PS(0.5)	0.536	0.092	0.063	0.055	-0.007	0.517
FedCvg(0.0001)+LD+PS(0.7)	0.538	0.093	0.063	0.056	-0.005	0.525
FedCvg(0.01)+LD+PS(0.7)	0.537	0.094	0.065	0.057	-0.008	0.517
FedCvg-Ratio(0.5)	0.540	0.084	0.060	0.052	-0.013	0.527
FedCvg-Ratio(0.9)	0.540	0.084	0.060	0.052	-0.013	0.527
FedCvg-Ratio(0.5)+LD	0.540	0.088	0.060	0.053	-0.010	0.525
FedCvg-Ratio(0.9)+LD	0.540	0.088	0.060	0.053	-0.010	0.525
FedCvg-Ratio(0.5)+PS(0.5)	0.540	0.082	0.058	0.050	-0.013	0.527
FedCvg-Ratio(0.9)+PS(0.5)	0.540	0.085	0.058	0.050	-0.014	0.528
FedCvg-Ratio(0.5)+PS(0.7)	0.540	0.080	0.057	0.050	-0.014	0.527
FedCvg-Ratio(0.9)+PS(0.7)	0.541	0.085	0.057	0.049	-0.014	0.530
FedCvg-Ratio(0.5)+LD+PS(0.5)	0.540	0.085	0.059	0.051	-0.014	0.526
FedCvg-Ratio(0.9)+LD+PS(0.5)	0.540	0.087	0.062	0.054	-0.006	0.526
FedCvg-Ratio(0.5)+LD+PS(0.7)	0.563	0.086	0.063	0.058	-0.027	**0.597**
FedCvg-Ratio(0.9)+LD+PS(0.7)	0.541	0.084	0.058	0.050	-0.014	0.528

Table B.5: Compas dataset: Performance at $\alpha = 5000.0$ (best values across learning rates, averaged over 5 seeds)

Method	Acc	EOD	SPD	AOD	AccDiff	Prec
FedAvg	0.522	0.053	0.044	0.037	-0.046	0.497
FedAvg+LD	0.522	0.051	0.042	0.036	-0.044	0.497
FedAvg+PS(0.5)	0.524	0.056	0.044	0.038	-0.046	0.499
FedAvg+PS(0.7)	0.524	0.056	0.044	0.038	-0.046	0.499
FedAvg+LD+PS(0.5)	0.522	0.054	0.044	0.038	-0.044	0.498
FedAvg+LD+PS(0.7)	0.524	0.055	0.044	0.038	-0.045	0.498
FairFed(EOD)	0.522	0.055	0.042	0.037	-0.046	0.497
FairFed(SPD)	0.523	0.054	0.043	0.037	-0.046	0.498
FairFed(AccDiff)	0.521	0.055	0.045	0.039	-0.046	0.496
FairFed(EOD)+LD	0.523	0.053	0.042	0.036	-0.046	0.498
FairFed(SPD)+LD	0.523	0.054	0.043	0.037	-0.046	0.498
FairFed(AccDiff)+LD	0.521	0.054	0.045	0.039	-0.045	0.496
FairFed(EOD)+PS(0.5)	0.523	**0.049**	**0.039**	**0.033**	-0.045	0.497

Continued on next page

Table B.5 – continued from previous page

Method	Acc	EOD	SPD	AOD	AccDiff	Prec
FairFed(SPD)+PS(0.5)	0.524	0.054	0.043	0.037	-0.046	0.499
FairFed(AccDiff)+PS(0.5)	0.523	0.056	0.043	0.038	-0.045	0.498
FairFed(EOD)+PS(0.7)	0.548	0.052	0.053	0.040	-0.051	0.529
FairFed(SPD)+PS(0.7)	0.524	0.054	0.044	0.038	-0.047	0.499
FairFed(AccDiff)+PS(0.7)	0.523	0.053	0.042	0.037	-0.046	0.498
FairFed(EOD)+LD+PS(0.5)	0.524	0.051	0.040	0.034	-0.047	0.499
FairFed(SPD)+LD+PS(0.5)	0.523	0.051	0.041	0.035	-0.048	0.498
FairFed(AccDiff)+LD+PS(0.5)	0.523	0.056	0.044	0.039	-0.045	0.498
FairFed(EOD)+LD+PS(0.7)	0.523	0.052	0.040	0.034	-0.046	0.498
FairFed(SPD)+LD+PS(0.7)	0.523	0.051	0.041	0.035	-0.048	0.498
FairFed(AccDiff)+LD+PS(0.7)	0.523	0.055	0.043	0.038	-0.047	0.498
FedCvg(0.0001)	**0.571**	0.065	0.074	0.058	-0.051	0.554
FedCvg(0.01)	0.557	0.065	0.063	0.053	-0.042	0.532
FedCvg(0.0001)+LD	0.570	0.064	0.074	0.058	-0.051	0.551
FedCvg(0.01)+LD	0.556	0.064	0.062	0.052	-0.042	0.529
FedCvg(0.0001)+PS(0.5)	0.557	0.070	0.063	0.054	-0.042	0.534
FedCvg(0.01)+PS(0.5)	0.557	0.069	0.062	0.053	-0.042	0.534
FedCvg(0.0001)+PS(0.7)	0.559	0.068	0.062	0.053	-0.043	0.539
FedCvg(0.01)+PS(0.7)	0.558	0.068	0.062	0.053	-0.044	0.536
FedCvg(0.0001)+LD+PS(0.5)	0.558	0.061	0.059	0.049	-0.045	0.534
FedCvg(0.01)+LD+PS(0.5)	0.557	0.062	0.060	0.050	-0.042	0.530
FedCvg(0.0001)+LD+PS(0.7)	0.558	0.061	0.058	0.048	-0.044	0.532
FedCvg(0.01)+LD+PS(0.7)	0.558	0.062	0.059	0.049	-0.044	0.533
FedCvg-Ratio(0.5)	**0.571**	0.064	0.072	0.057	**-0.052**	**0.556**
FedCvg-Ratio(0.9)	**0.571**	0.064	0.072	0.057	**-0.052**	**0.556**
FedCvg-Ratio(0.5)+LD	0.570	0.066	0.074	0.059	-0.049	0.553
FedCvg-Ratio(0.9)+LD	0.570	0.066	0.074	0.059	-0.049	0.553
FedCvg-Ratio(0.5)+PS(0.5)	0.558	0.064	0.060	0.050	-0.044	0.537
FedCvg-Ratio(0.9)+PS(0.5)	0.558	0.063	0.060	0.050	-0.045	0.536
FedCvg-Ratio(0.5)+PS(0.7)	0.558	0.063	0.060	0.050	-0.045	0.537
FedCvg-Ratio(0.9)+PS(0.7)	0.558	0.065	0.061	0.051	-0.044	0.535
FedCvg-Ratio(0.5)+LD+PS(0.5)	0.557	0.062	0.060	0.050	-0.044	0.532
FedCvg-Ratio(0.9)+LD+PS(0.5)	0.558	0.061	0.060	0.050	-0.045	0.533
FedCvg-Ratio(0.5)+LD+PS(0.7)	0.558	0.062	0.059	0.049	-0.044	0.531
FedCvg-Ratio(0.9)+LD+PS(0.7)	0.558	0.063	0.060	0.050	-0.043	0.532

Appendix C

Coverage-based Partitioning Results: Adult Dataset

Table C.1: Adult dataset: Performance at coverage=1999, partition=diff_size (best EOD across learning rates, averaged over 5 seeds)

Method	Acc	EOD	SPD	AOD	AccDiff	Prec
FairFed(EOD)	0.796	0.054	0.037	0.032	0.166	0.788
FairFed(SPD)	0.797	0.060	0.040	0.035	0.163	0.781
FairFed(AccDiff)	0.797	0.059	0.040	0.035	0.163	0.783
FairFed(EOD)+LD	0.793	0.034	**0.030**	0.021	0.170	0.790
FairFed(SPD)+LD	0.795	0.036	0.033	0.022	0.168	0.783
FairFed(AccDiff)+LD	0.805	0.037	0.049	0.020	0.157	0.785
FairFed(EOD)+PS(0.5)	0.795	0.054	0.037	0.032	0.166	0.790
FairFed(SPD)+PS(0.5)	0.795	0.055	0.038	0.032	0.165	0.782
FairFed(AccDiff)+PS(0.5)	0.796	0.057	0.039	0.034	0.164	0.784
FairFed(EOD)+PS(0.7)	0.795	0.054	0.038	0.032	0.166	0.782
FairFed(SPD)+PS(0.7)	0.794	0.053	0.037	0.031	0.166	0.778
FairFed(AccDiff)+PS(0.7)	0.796	0.057	0.040	0.034	0.164	0.778
FairFed(EOD)+LD+PS(0.5)	0.793	0.036	0.030	0.022	0.170	**0.792**
FairFed(SPD)+LD+PS(0.5)	0.804	0.032	0.047	0.017	0.158	0.786
FairFed(AccDiff)+LD+PS(0.5)	0.793	0.035	0.031	0.021	0.169	0.786
FairFed(EOD)+LD+PS(0.7)	0.794	0.034	0.031	0.021	0.170	0.788
FairFed(SPD)+LD+PS(0.7)	0.794	0.033	0.031	0.020	0.169	0.788
FairFed(AccDiff)+LD+PS(0.7)	0.794	0.035	0.031	0.021	0.169	0.787
FedAvg	0.797	0.059	0.040	0.035	0.163	0.782
FedAvg+LD	0.795	0.037	0.033	0.023	0.168	0.784
FedAvg+PS(0.5)	0.796	0.056	0.038	0.033	0.165	0.781

Continued on next page

Table C.1 – continued from previous page

Method	Acc	EOD	SPD	AOD	AccDiff	Prec
FedAvg+PS(0.7)	0.796	0.056	0.038	0.033	0.164	0.783
FedAvg+LD+PS(0.5)	0.794	0.035	0.031	0.021	0.169	0.791
FedAvg+LD+PS(0.7)	0.794	0.033	0.031	0.021	0.169	0.790
FedCvg(0.0001)	0.797	0.062	0.037	0.035	0.170	0.773
FedCvg(0.001)	0.796	0.060	0.035	0.034	0.172	0.785
FedCvg(0.0001)+LD	0.823	0.033	0.069	0.020	**0.143**	0.739
FedCvg(0.001)+LD	0.822	0.040	0.069	0.023	**0.143**	0.745
FedCvg(0.0001)+PS(0.5)	0.796	0.060	0.036	0.034	0.171	0.777
FedCvg(0.001)+PS(0.5)	0.795	0.058	0.034	0.033	0.173	0.790
FedCvg(0.0001)+PS(0.7)	0.796	0.059	0.036	0.033	0.172	0.778
FedCvg(0.001)+PS(0.7)	0.795	0.058	0.034	0.032	0.173	0.789
FedCvg(0.0001)+LD+PS(0.5)	0.822	0.034	0.068	0.020	0.144	0.737
FedCvg(0.001)+LD+PS(0.5)	0.808	0.036	0.051	0.023	0.159	0.763
FedCvg(0.0001)+LD+PS(0.7)	0.809	0.032	0.053	0.022	0.158	0.756
FedCvg(0.001)+LD+PS(0.7)	0.822	0.035	0.066	0.020	0.145	0.744
FedCvg-Ratio(0.5)	0.818	0.034	0.075	0.029	0.154	0.752
FedCvg-Ratio(0.9)	0.818	0.034	0.075	0.029	0.154	0.752
FedCvg-Ratio(0.5)+LD	0.827	0.012	0.077	0.011	0.147	0.782
FedCvg-Ratio(0.9)+LD	0.827	0.012	0.077	0.011	0.147	0.782
FedCvg-Ratio(0.5)+PS(0.5)	**0.829**	0.028	0.098	0.032	0.145	0.763
FedCvg-Ratio(0.9)+PS(0.5)	**0.829**	0.032	0.098	0.034	0.144	0.764
FedCvg-Ratio(0.5)+PS(0.7)	0.828	0.029	0.097	0.032	0.145	0.755
FedCvg-Ratio(0.9)+PS(0.7)	0.828	0.029	0.098	0.033	0.145	0.754
FedCvg-Ratio(0.5)+LD+PS(0.5)	0.815	0.009	0.060	0.010	0.160	0.787
FedCvg-Ratio(0.9)+LD+PS(0.5)	0.815	0.011	0.058	**0.010**	0.160	0.786
FedCvg-Ratio(0.5)+LD+PS(0.7)	0.816	0.009	0.060	0.012	0.158	0.784
FedCvg-Ratio(0.9)+LD+PS(0.7)	0.828	**0.008**	0.080	0.013	0.146	0.782

Table C.2: Adult dataset: Performance at coverage=1999, partition=same_size (best EOD across learning rates, averaged over 5 seeds)

Method	Acc	EOD	SPD	AOD	AccDiff	Prec
FairFed(EOD)	0.798	0.049	0.032	0.028	0.176	0.773
FairFed(SPD)	0.798	0.049	0.032	0.029	0.176	0.783
FairFed(AccDiff)	0.799	0.051	0.033	0.030	0.175	0.777
FairFed(EOD)+LD	0.820	0.035	0.053	0.015	0.156	0.775
FairFed(SPD)+LD	0.820	0.029	0.056	**0.011**	0.157	0.773
FairFed(AccDiff)+LD	0.809	0.038	0.041	0.019	0.167	0.782
FairFed(EOD)+PS(0.5)	0.797	0.046	0.031	0.027	0.177	0.783
FairFed(SPD)+PS(0.5)	0.797	0.048	0.032	0.028	0.177	0.775

Continued on next page

Table C.2 – continued from previous page

Method	Acc	EOD	SPD	AOD	AccDiff	Prec
FairFed(AccDiff)+PS(0.5)	0.798	0.050	0.032	0.029	0.176	0.777
FairFed(EOD)+PS(0.7)	0.797	0.046	0.031	0.027	0.177	0.780
FairFed(SPD)+PS(0.7)	0.797	0.046	0.031	0.027	0.177	0.776
FairFed(AccDiff)+PS(0.7)	0.798	0.049	0.032	0.028	0.176	0.776
FairFed(EOD)+LD+PS(0.5)	0.808	0.033	0.040	0.016	0.168	0.784
FairFed(SPD)+LD+PS(0.5)	0.807	0.032	0.041	0.016	0.168	0.783
FairFed(AccDiff)+LD+PS(0.5)	0.808	0.038	0.041	0.019	0.168	0.784
FairFed(EOD)+LD+PS(0.7)	0.807	0.034	0.039	0.017	0.168	0.785
FairFed(SPD)+LD+PS(0.7)	0.807	0.032	0.040	0.015	0.168	0.781
FairFed(AccDiff)+LD+PS(0.7)	0.808	0.037	0.041	0.018	0.168	0.781
FedAvg	0.798	0.050	0.032	0.029	0.176	0.782
FedAvg+LD	0.808	0.036	0.041	0.018	0.169	0.778
FedAvg+PS(0.5)	0.798	0.049	0.032	0.028	0.176	0.777
FedAvg+PS(0.7)	0.798	0.048	0.032	0.028	0.176	0.777
FedAvg+LD+PS(0.5)	0.808	0.031	0.041	0.015	0.168	0.778
FedAvg+LD+PS(0.7)	0.808	0.033	0.041	0.016	0.168	0.784
FedCvg(0.0001)	0.795	0.045	0.029	0.024	0.176	0.802
FedCvg(0.001)	0.794	0.043	0.027	0.024	0.179	**0.809**
FedCvg(0.0001)+LD	0.799	0.029	0.035	0.018	0.174	0.808
FedCvg(0.001)+LD	0.792	0.032	**0.022**	0.017	0.181	0.801
FedCvg(0.0001)+PS(0.5)	0.795	0.042	0.029	0.023	0.176	0.798
FedCvg(0.001)+PS(0.5)	0.794	0.039	0.027	0.022	0.178	0.797
FedCvg(0.0001)+PS(0.7)	0.794	0.041	0.028	0.023	0.177	0.795
FedCvg(0.001)+PS(0.7)	0.793	0.038	0.027	0.021	0.178	0.795
FedCvg(0.0001)+LD+PS(0.5)	0.793	0.027	0.023	0.015	0.179	0.796
FedCvg(0.001)+LD+PS(0.5)	0.792	0.028	**0.022**	0.015	0.181	0.797
FedCvg(0.0001)+LD+PS(0.7)	0.793	0.027	0.023	0.014	0.180	0.794
FedCvg(0.001)+LD+PS(0.7)	0.800	0.027	0.036	0.017	0.173	0.787
FedCvg-Ratio(0.5)	0.817	0.031	0.074	0.020	0.161	0.786
FedCvg-Ratio(0.9)	0.817	0.031	0.074	0.020	0.161	0.786
FedCvg-Ratio(0.5)+LD	0.826	**0.012**	0.074	0.015	0.149	0.778
FedCvg-Ratio(0.9)+LD	0.826	**0.012**	0.074	0.015	0.149	0.778
FedCvg-Ratio(0.5)+PS(0.5)	0.819	0.038	0.077	0.024	0.158	0.785
FedCvg-Ratio(0.9)+PS(0.5)	0.819	0.038	0.077	0.024	0.158	0.786
FedCvg-Ratio(0.5)+PS(0.7)	0.819	0.039	0.077	0.024	0.158	0.785
FedCvg-Ratio(0.9)+PS(0.7)	0.819	0.038	0.077	0.024	0.158	0.785
FedCvg-Ratio(0.5)+LD+PS(0.5)	**0.838**	0.014	0.094	0.020	**0.136**	0.750
FedCvg-Ratio(0.9)+LD+PS(0.5)	0.827	0.013	0.075	0.016	0.148	0.776
FedCvg-Ratio(0.5)+LD+PS(0.7)	**0.838**	0.013	0.094	0.020	**0.137**	0.749
FedCvg-Ratio(0.9)+LD+PS(0.7)	**0.838**	0.013	0.094	0.020	**0.136**	0.748

Table C.3: Adult dataset: Performance at coverage=2570, partition=diff_size (best EOD across learning rates, averaged over 5 seeds)

Method	Acc	EOD	SPD	AOD	AccDiff	Prec
FairFed(EOD)	0.797	0.063	0.039	0.036	0.167	0.787
FairFed(SPD)	0.799	0.070	0.043	0.040	0.164	0.781
FairFed(AccDiff)	0.798	0.068	0.040	0.038	0.165	0.789
FairFed(EOD)+LD	0.795	0.041	0.031	0.024	0.171	0.792
FairFed(SPD)+LD	0.797	0.048	0.036	0.028	0.168	0.782
FairFed(AccDiff)+LD	0.796	0.044	0.033	0.026	0.170	0.794
FairFed(EOD)+PS(0.5)	0.796	0.069	0.039	0.039	0.166	0.787
FairFed(SPD)+PS(0.5)	0.797	0.069	0.040	0.039	0.165	0.785
FairFed(AccDiff)+PS(0.5)	0.797	0.068	0.039	0.039	0.166	0.786
FairFed(EOD)+PS(0.7)	0.796	0.067	0.039	0.038	0.167	0.789
FairFed(SPD)+PS(0.7)	0.797	0.074	0.042	0.042	0.165	0.781
FairFed(AccDiff)+PS(0.7)	0.797	0.069	0.040	0.039	0.166	0.783
FairFed(EOD)+LD+PS(0.5)	0.793	0.040	0.030	0.023	0.172	0.790
FairFed(SPD)+LD+PS(0.5)	0.795	0.047	0.033	0.027	0.170	0.787
FairFed(AccDiff)+LD+PS(0.5)	0.794	0.048	0.032	0.028	0.170	0.788
FairFed(EOD)+LD+PS(0.7)	0.793	0.038	**0.029**	0.022	0.173	0.794
FairFed(SPD)+LD+PS(0.7)	0.795	0.044	0.034	0.026	0.170	0.777
FairFed(AccDiff)+LD+PS(0.7)	0.793	0.042	0.030	0.024	0.172	0.794
FedAvg	0.798	0.071	0.042	0.040	0.164	0.782
FedAvg+LD	0.796	0.049	0.035	0.028	0.168	0.787
FedAvg+PS(0.5)	0.797	0.070	0.040	0.039	0.166	0.785
FedAvg+PS(0.7)	0.797	0.067	0.039	0.038	0.166	0.784
FedAvg+LD+PS(0.5)	0.794	0.048	0.033	0.028	0.170	0.783
FedAvg+LD+PS(0.7)	0.794	0.046	0.032	0.027	0.171	0.784
FedCvg(0.0001)	0.801	0.056	0.039	0.032	0.163	0.796
FedCvg(0.001)	0.798	0.049	0.033	0.028	0.167	0.808
FedCvg(0.0001)+LD	0.811	0.039	0.045	0.018	0.154	0.787
FedCvg(0.001)+LD	0.808	0.028	0.041	0.013	0.157	0.794
FedCvg(0.0001)+PS(0.5)	0.800	0.055	0.037	0.030	0.165	0.800
FedCvg(0.001)+PS(0.5)	0.798	0.047	0.033	0.028	0.167	0.812
FedCvg(0.0001)+PS(0.7)	0.799	0.054	0.037	0.030	0.165	0.799
FedCvg(0.001)+PS(0.7)	0.797	0.048	0.033	0.028	0.167	**0.813**
FedCvg(0.0001)+LD+PS(0.5)	0.809	0.038	0.044	0.018	0.156	0.787
FedCvg(0.001)+LD+PS(0.5)	0.815	0.024	0.054	0.015	0.152	0.783
FedCvg(0.0001)+LD+PS(0.7)	0.809	0.035	0.043	0.016	0.156	0.787
FedCvg(0.001)+LD+PS(0.7)	0.815	0.025	0.053	0.014	0.152	0.783
FedCvg-Ratio(0.5)	0.821	0.030	0.076	0.029	0.149	0.780
FedCvg-Ratio(0.9)	0.821	0.030	0.076	0.029	0.149	0.780
FedCvg-Ratio(0.5)+LD	**0.825**	**0.010**	0.075	0.013	**0.141**	0.775

Continued on next page

Table C.3 – continued from previous page

Method	Acc	EOD	SPD	AOD	AccDiff	Prec
FedCvg-Ratio(0.9)+LD	**0.825**	**0.010**	0.075	0.013	**0.141**	0.775
FedCvg-Ratio(0.5)+PS(0.5)	0.822	0.028	0.075	0.028	0.148	0.779
FedCvg-Ratio(0.9)+PS(0.5)	0.822	0.033	0.077	0.030	0.148	0.780
FedCvg-Ratio(0.5)+PS(0.7)	0.821	0.028	0.076	0.028	0.148	0.779
FedCvg-Ratio(0.9)+PS(0.7)	0.822	0.031	0.077	0.029	0.147	0.779
FedCvg-Ratio(0.5)+LD+PS(0.5)	0.824	**0.010**	0.072	0.013	0.142	0.775
FedCvg-Ratio(0.9)+LD+PS(0.5)	0.824	0.011	0.073	0.012	0.142	0.774
FedCvg-Ratio(0.5)+LD+PS(0.7)	0.824	0.014	0.073	**0.011**	0.142	0.776
FedCvg-Ratio(0.9)+LD+PS(0.7)	0.810	0.014	0.052	0.014	0.155	0.795

Table C.4: Adult dataset: Performance at coverage=2570, partition=same_size (best EOD across learning rates, averaged over 5 seeds)

Method	Acc	EOD	SPD	AOD	AccDiff	Prec
FairFed(EOD)	0.794	0.052	0.031	0.030	0.176	0.776
FairFed(SPD)	0.794	0.053	0.032	0.031	0.176	0.775
FairFed(AccDiff)	0.794	0.051	0.031	0.029	0.176	0.778
FairFed(EOD)+LD	0.800	0.033	0.039	0.022	0.173	0.790
FairFed(SPD)+LD	0.801	0.034	0.040	0.023	0.172	0.789
FairFed(AccDiff)+LD	0.792	0.036	0.025	0.021	0.179	0.792
FairFed(EOD)+PS(0.5)	0.793	0.053	0.030	0.030	0.177	0.776
FairFed(SPD)+PS(0.5)	0.794	0.052	0.030	0.030	0.176	0.782
FairFed(AccDiff)+PS(0.5)	0.794	0.049	0.030	0.028	0.177	0.784
FairFed(EOD)+PS(0.7)	0.793	0.052	0.029	0.030	0.177	0.778
FairFed(SPD)+PS(0.7)	0.793	0.054	0.030	0.031	0.176	0.782
FairFed(AccDiff)+PS(0.7)	0.793	0.053	0.029	0.030	0.177	0.781
FairFed(EOD)+LD+PS(0.5)	0.800	0.033	0.038	0.021	0.173	0.794
FairFed(SPD)+LD+PS(0.5)	0.800	0.032	0.039	0.022	0.173	0.790
FairFed(AccDiff)+LD+PS(0.5)	0.800	0.038	0.039	0.025	0.172	0.795
FairFed(EOD)+LD+PS(0.7)	0.800	0.034	0.037	0.022	0.173	0.792
FairFed(SPD)+LD+PS(0.7)	0.800	0.033	0.038	0.022	0.173	0.796
FairFed(AccDiff)+LD+PS(0.7)	0.800	0.035	0.038	0.023	0.173	0.798
FedAvg	0.794	0.050	0.031	0.029	0.175	0.774
FedAvg+LD	0.793	0.035	0.026	0.020	0.178	0.789
FedAvg+PS(0.5)	0.793	0.053	0.030	0.030	0.176	0.780
FedAvg+PS(0.7)	0.793	0.053	0.029	0.030	0.177	0.780
FedAvg+LD+PS(0.5)	0.800	0.035	0.039	0.023	0.173	0.790
FedAvg+LD+PS(0.7)	0.800	0.034	0.038	0.023	0.173	0.791
FedCvg(0.0001)	0.793	0.039	0.027	0.022	0.175	0.805
FedCvg(0.001)	0.790	0.034	0.023	0.020	0.179	0.812

Continued on next page

Table C.4 – continued from previous page

Method	Acc	EOD	SPD	AOD	AccDiff	Prec
FedCvg(0.0001)+LD	0.791	0.031	0.022	0.016	0.179	0.808
FedCvg(0.001)+LD	0.789	0.021	0.018	0.012	0.182	0.828
FedCvg(0.0001)+PS(0.5)	0.792	0.038	0.026	0.021	0.176	0.807
FedCvg(0.001)+PS(0.5)	0.790	0.034	0.023	0.019	0.179	0.810
FedCvg(0.0001)+PS(0.7)	0.792	0.038	0.026	0.021	0.177	0.811
FedCvg(0.001)+PS(0.7)	0.790	0.033	0.023	0.019	0.180	0.815
FedCvg(0.0001)+LD+PS(0.5)	0.791	0.028	0.020	0.015	0.180	0.813
FedCvg(0.001)+LD+PS(0.5)	0.789	0.021	0.018	0.012	0.181	0.830
FedCvg(0.0001)+LD+PS(0.7)	0.798	0.025	0.034	0.016	0.174	0.813
FedCvg(0.001)+LD+PS(0.7)	0.789	0.020	**0.017**	**0.011**	0.182	**0.832**
FedCvg-Ratio(0.5)	0.803	0.031	0.045	0.023	0.169	0.808
FedCvg-Ratio(0.9)	0.803	0.031	0.045	0.023	0.169	0.808
FedCvg-Ratio(0.5)+LD	0.816	**0.013**	0.060	0.014	0.158	0.799
FedCvg-Ratio(0.9)+LD	0.816	**0.013**	0.060	0.014	0.158	0.799
FedCvg-Ratio(0.5)+PS(0.5)	0.803	0.031	0.047	0.023	0.168	0.798
FedCvg-Ratio(0.9)+PS(0.5)	0.792	0.031	0.024	0.018	0.177	0.805
FedCvg-Ratio(0.5)+PS(0.7)	0.792	0.032	0.025	0.018	0.177	0.810
FedCvg-Ratio(0.9)+PS(0.7)	0.792	0.031	0.024	0.018	0.177	0.811
FedCvg-Ratio(0.5)+LD+PS(0.5)	0.824	0.014	0.073	0.016	0.151	0.795
FedCvg-Ratio(0.9)+LD+PS(0.5)	0.824	0.014	0.073	0.017	0.150	0.790
FedCvg-Ratio(0.5)+LD+PS(0.7)	**0.831**	0.014	0.086	0.019	**0.144**	0.790
FedCvg-Ratio(0.9)+LD+PS(0.7)	0.824	**0.013**	0.073	0.015	0.152	0.796

Table C.5: Adult dataset: Performance at coverage=4497, partition=diff_size (best EOD across learning rates, averaged over 5 seeds)

Method	Acc	EOD	SPD	AOD	AccDiff	Prec
FairFed(EOD)	0.812	0.098	0.096	0.065	0.140	0.765
FairFed(SPD)	0.802	0.100	0.076	0.062	0.148	0.780
FairFed(AccDiff)	0.812	0.101	0.097	0.067	0.140	0.766
FairFed(EOD)+LD	0.819	0.054	0.083	0.032	0.131	0.762
FairFed(SPD)+LD	0.819	0.048	0.097	0.037	0.132	0.759
FairFed(AccDiff)+LD	0.819	0.049	0.100	0.039	0.132	0.755
FairFed(EOD)+PS(0.5)	0.800	0.088	0.071	0.055	0.152	0.785
FairFed(SPD)+PS(0.5)	0.799	0.084	0.070	0.053	0.153	0.779
FairFed(AccDiff)+PS(0.5)	0.810	0.092	0.092	0.062	0.143	0.770
FairFed(EOD)+PS(0.7)	0.810	0.088	0.092	0.060	0.144	0.775
FairFed(SPD)+PS(0.7)	0.810	0.088	0.092	0.060	0.144	0.773
FairFed(AccDiff)+PS(0.7)	0.800	0.087	0.071	0.054	0.152	0.788
FairFed(EOD)+LD+PS(0.5)	0.818	0.055	0.081	0.030	0.132	0.763
FairFed(SPD)+LD+PS(0.5)	0.811	0.063	0.076	0.036	0.139	0.774

Continued on next page

Table C.5 – continued from previous page

Method	Acc	EOD	SPD	AOD	AccDiff	Prec
FairFed(AccDiff)+LD+PS(0.5)	0.819	0.051	0.083	0.030	0.132	0.763
FairFed(EOD)+LD+PS(0.7)	0.817	0.059	0.091	0.039	0.135	0.766
FairFed(SPD)+LD+PS(0.7)	0.804	0.056	0.074	0.040	0.149	0.781
FairFed(AccDiff)+LD+PS(0.7)	0.818	0.057	0.091	0.037	0.135	0.767
FedAvg	0.812	0.097	0.096	0.065	0.141	0.765
FedAvg+LD	0.820	0.051	0.086	0.030	0.130	0.756
FedAvg+PS(0.5)	0.810	0.089	0.092	0.061	0.143	0.768
FedAvg+PS(0.7)	0.810	0.086	0.091	0.059	0.144	0.773
FedAvg+LD+PS(0.5)	0.819	0.051	0.082	0.031	0.132	0.761
FedAvg+LD+PS(0.7)	0.818	0.054	0.080	0.031	0.132	0.763
FedCvg(0.0001)	0.809	0.073	0.094	0.054	0.147	0.770
FedCvg(0.001)	0.807	0.063	0.090	0.049	0.150	0.776
FedCvg(0.0001)+LD	0.804	0.037	0.074	0.031	0.151	0.778
FedCvg(0.001)+LD	0.802	0.025	0.069	0.024	0.154	0.786
FedCvg(0.0001)+PS(0.5)	0.808	0.069	0.092	0.052	0.148	0.771
FedCvg(0.001)+PS(0.5)	0.807	0.064	0.089	0.049	0.150	0.777
FedCvg(0.0001)+PS(0.7)	0.807	0.066	0.090	0.050	0.149	0.774
FedCvg(0.001)+PS(0.7)	0.806	0.063	0.089	0.049	0.150	0.774
FedCvg(0.0001)+LD+PS(0.5)	0.802	0.037	0.072	0.030	0.153	0.780
FedCvg(0.001)+LD+PS(0.5)	0.801	0.030	0.069	0.026	0.154	**0.791**
FedCvg(0.0001)+LD+PS(0.7)	0.802	0.031	0.070	0.027	0.154	0.779
FedCvg(0.001)+LD+PS(0.7)	0.801	0.029	0.069	0.026	0.154	0.789
FedCvg-Ratio(0.5)	**0.846**	0.029	0.143	0.034	**0.118**	0.737
FedCvg-Ratio(0.9)	**0.846**	0.029	0.143	0.034	**0.118**	0.737
FedCvg-Ratio(0.5)+LD	0.802	**0.012**	**0.062**	**0.013**	0.153	0.784
FedCvg-Ratio(0.9)+LD	0.802	**0.012**	**0.062**	**0.013**	0.153	0.784
FedCvg-Ratio(0.5)+PS(0.5)	**0.846**	0.025	0.142	0.033	**0.119**	0.738
FedCvg-Ratio(0.9)+PS(0.5)	**0.846**	0.028	0.143	0.033	**0.118**	0.739
FedCvg-Ratio(0.5)+PS(0.7)	**0.846**	0.026	0.142	0.032	**0.118**	0.738
FedCvg-Ratio(0.9)+PS(0.7)	**0.846**	0.025	0.142	0.032	**0.118**	0.737
FedCvg-Ratio(0.5)+LD+PS(0.5)	0.802	0.012	0.062	**0.014**	0.153	0.783
FedCvg-Ratio(0.9)+LD+PS(0.5)	0.802	0.014	**0.062**	0.015	0.152	0.785
FedCvg-Ratio(0.5)+LD+PS(0.7)	0.801	**0.012**	**0.062**	**0.014**	0.153	0.781
FedCvg-Ratio(0.9)+LD+PS(0.7)	0.801	**0.012**	**0.062**	**0.013**	0.153	0.781

Table C.6: Adult dataset: Performance at coverage=4497, partition=same_size (best EOD across learning rates, averaged over 5 seeds)

Method	Acc	EOD	SPD	AOD	AccDiff	Prec
FairFed(EOD)	0.780	0.058	0.032	0.032	0.178	0.822
FairFed(SPD)	0.782	0.064	0.037	0.036	0.175	0.812
FairFed(AccDiff)	0.783	0.066	0.038	0.037	0.173	0.813
FairFed(EOD)+LD	0.777	0.037	0.025	0.021	0.182	0.829
FairFed(SPD)+LD	0.792	0.042	0.045	0.018	0.168	0.813
FairFed(AccDiff)+LD	0.780	0.042	0.030	0.024	0.178	0.820
FairFed(EOD)+PS(0.5)	0.780	0.057	0.031	0.032	0.178	0.824
FairFed(SPD)+PS(0.5)	0.782	0.062	0.036	0.035	0.175	0.813
FairFed(AccDiff)+PS(0.5)	0.782	0.062	0.036	0.035	0.175	0.814
FairFed(EOD)+PS(0.7)	0.779	0.055	0.030	0.031	0.179	0.831
FairFed(SPD)+PS(0.7)	0.781	0.061	0.034	0.034	0.176	0.821
FairFed(AccDiff)+PS(0.7)	0.781	0.062	0.035	0.034	0.176	0.816
FairFed(EOD)+LD+PS(0.5)	0.777	0.036	0.024	0.020	0.182	0.835
FairFed(SPD)+LD+PS(0.5)	0.779	0.039	0.028	0.022	0.179	0.821
FairFed(AccDiff)+LD+PS(0.5)	0.779	0.040	0.028	0.023	0.179	0.823
FairFed(EOD)+LD+PS(0.7)	0.776	0.034	**0.023**	0.019	0.183	**0.836**
FairFed(SPD)+LD+PS(0.7)	0.778	0.039	0.027	0.022	0.180	0.825
FairFed(AccDiff)+LD+PS(0.7)	0.779	0.037	0.026	0.021	0.180	0.826
FedAvg	0.786	0.071	0.043	0.040	0.170	0.803
FedAvg+LD	0.796	0.047	0.051	0.021	0.164	0.799
FedAvg+PS(0.5)	0.782	0.062	0.037	0.035	0.175	0.811
FedAvg+PS(0.7)	0.781	0.063	0.035	0.035	0.176	0.815
FedAvg+LD+PS(0.5)	0.780	0.042	0.030	0.024	0.178	0.822
FedAvg+LD+PS(0.7)	0.779	0.040	0.027	0.022	0.180	0.826
FedCvg(0.0001)	0.800	0.055	0.062	0.038	0.156	0.797
FedCvg(0.001)	0.794	0.038	0.051	0.028	0.163	0.801
FedCvg(0.0001)+LD	0.785	0.035	0.031	0.020	0.169	0.819
FedCvg(0.001)+LD	0.786	0.021	0.032	0.014	0.170	0.824
FedCvg(0.0001)+PS(0.5)	0.798	0.051	0.059	0.036	0.158	0.799
FedCvg(0.001)+PS(0.5)	0.794	0.041	0.052	0.029	0.162	0.801
FedCvg(0.0001)+PS(0.7)	0.797	0.050	0.057	0.035	0.159	0.800
FedCvg(0.001)+PS(0.7)	0.794	0.037	0.051	0.027	0.163	0.799
FedCvg(0.0001)+LD+PS(0.5)	0.782	0.030	0.026	0.018	0.172	0.830
FedCvg(0.001)+LD+PS(0.5)	0.787	0.023	0.033	0.016	0.169	0.825
FedCvg(0.0001)+LD+PS(0.7)	0.781	0.031	0.025	0.018	0.173	0.832
FedCvg(0.001)+LD+PS(0.7)	0.787	0.024	0.033	0.016	0.170	0.826
FedCvg-Ratio(0.5)	0.820	0.026	0.095	0.031	0.140	0.783
FedCvg-Ratio(0.9)	0.820	0.026	0.095	0.031	0.140	0.783
FedCvg-Ratio(0.5)+LD	0.816	0.018	0.072	**0.008**	0.145	0.787

Continued on next page

Table C.6 – continued from previous page

Method	Acc	EOD	SPD	AOD	AccDiff	Prec
FedCvg-Ratio(0.9)+LD	0.816	0.018	0.072	**0.008**	0.145	0.787
FedCvg-Ratio(0.5)+PS(0.5)	0.821	0.029	0.097	0.033	0.140	0.779
FedCvg-Ratio(0.9)+PS(0.5)	0.821	0.029	0.097	0.033	0.140	0.780
FedCvg-Ratio(0.5)+PS(0.7)	0.821	0.029	0.098	0.033	0.140	0.778
FedCvg-Ratio(0.9)+PS(0.7)	0.821	0.026	0.098	0.032	0.140	0.779
FedCvg-Ratio(0.5)+LD+PS(0.5)	0.815	0.018	0.073	0.010	0.145	0.785
FedCvg-Ratio(0.9)+LD+PS(0.5)	**0.825**	0.017	0.090	0.014	**0.137**	0.773
FedCvg-Ratio(0.5)+LD+PS(0.7)	**0.825**	**0.016**	0.091	0.014	**0.137**	0.771
FedCvg-Ratio(0.9)+LD+PS(0.7)	**0.825**	0.017	0.090	0.013	**0.137**	0.771

Appendix D

Coverage-based Partitioning Results: COMPAS Dataset

Table D.1: Compas dataset: Performance at coverage=474, partition=diff_size (best EOD across learning rates, averaged over 5 seeds)

Method	Acc	EOD	SPD	AOD	AccDiff	Prec
FairFed(EOD)	**0.543**	**0.045**	0.048	0.038	**-0.059**	0.544
FairFed(SPD)	0.541	0.049	0.045	0.036	-0.055	0.544
FairFed(AccDiff)	0.542	0.049	0.045	0.036	-0.052	**0.550**
FairFed(EOD)+LD	0.538	0.046	0.042	0.033	-0.051	0.533
FairFed(SPD)+LD	0.539	0.051	0.044	0.037	-0.053	0.541
FairFed(AccDiff)+LD	**0.544**	0.051	0.049	0.040	-0.054	**0.549**
FairFed(EOD)+PS(0.5)	0.539	0.052	0.047	0.039	-0.055	0.539
FairFed(SPD)+PS(0.5)	0.539	0.054	0.043	0.035	-0.047	0.537
FairFed(AccDiff)+PS(0.5)	0.543	0.050	0.048	0.040	-0.053	0.538
FairFed(EOD)+PS(0.7)	0.542	0.049	0.046	0.037	-0.054	0.535
FairFed(SPD)+PS(0.7)	0.542	0.047	0.045	0.036	-0.052	0.532
FairFed(AccDiff)+PS(0.7)	**0.544**	0.057	0.048	0.040	-0.049	0.542
FairFed(EOD)+LD+PS(0.5)	0.537	0.045	0.038	**0.030**	-0.050	0.529
FairFed(SPD)+LD+PS(0.5)	0.542	0.047	0.046	0.037	-0.052	0.532
FairFed(AccDiff)+LD+PS(0.5)	**0.544**	0.051	0.050	0.041	-0.055	0.544
FairFed(EOD)+LD+PS(0.7)	0.543	0.046	0.047	0.038	-0.057	0.536
FairFed(SPD)+LD+PS(0.7)	0.537	0.055	0.044	0.037	-0.048	0.530
FairFed(AccDiff)+LD+PS(0.7)	0.539	0.048	0.043	0.035	-0.053	0.533
FedAvg	0.541	0.048	0.045	0.036	-0.052	0.533
FedAvg+LD	0.541	0.048	0.046	0.037	-0.051	0.531
FedAvg+PS(0.5)	0.541	0.048	0.045	0.036	-0.052	0.532

Continued on next page

Table D.1 – continued from previous page

Method	Acc	EOD	SPD	AOD	AccDiff	Prec
FedAvg+PS(0.7)	0.542	0.048	0.047	0.038	-0.053	0.533
FedAvg+LD+PS(0.5)	0.541	0.047	0.046	0.037	-0.052	0.531
FedAvg+LD+PS(0.7)	0.541	0.048	0.046	0.037	-0.052	0.531
FedCvg(0.0001)	0.510	0.075	0.038	0.038	-0.034	0.498
FedCvg(0.001)	0.512	0.074	0.042	0.040	-0.038	0.501
FedCvg(0.0001)+LD	0.510	0.076	0.038	0.038	-0.034	0.496
FedCvg(0.001)+LD	0.512	0.075	0.042	0.040	-0.037	0.501
FedCvg(0.0001)+PS(0.5)	0.510	0.073	0.037	0.036	-0.033	0.497
FedCvg(0.001)+PS(0.5)	0.511	0.072	0.037	0.036	-0.037	0.499
FedCvg(0.0001)+PS(0.7)	0.510	0.076	0.038	0.038	-0.034	0.498
FedCvg(0.001)+PS(0.7)	0.512	0.072	0.039	0.038	-0.037	0.500
FedCvg(0.0001)+LD+PS(0.5)	0.511	0.076	0.038	0.037	-0.033	0.499
FedCvg(0.001)+LD+PS(0.5)	0.510	0.075	0.037	0.036	-0.034	0.497
FedCvg(0.0001)+LD+PS(0.7)	0.511	0.077	0.039	0.038	-0.033	0.499
FedCvg(0.001)+LD+PS(0.7)	0.511	0.075	0.039	0.038	-0.036	0.500
FedCvg-Ratio(0.5)	0.532	0.070	0.044	0.044	-0.035	0.526
FedCvg-Ratio(0.9)	0.532	0.070	0.044	0.044	-0.035	0.526
FedCvg-Ratio(0.5)+LD	0.509	0.071	**0.036**	0.043	-0.034	0.491
FedCvg-Ratio(0.9)+LD	0.509	0.071	**0.036**	0.043	-0.034	0.491
FedCvg-Ratio(0.5)+PS(0.5)	0.508	0.069	**0.036**	0.043	-0.036	0.490
FedCvg-Ratio(0.9)+PS(0.5)	0.530	0.071	0.048	0.041	-0.036	0.521
FedCvg-Ratio(0.5)+PS(0.7)	0.525	0.069	0.043	0.045	-0.044	0.523
FedCvg-Ratio(0.9)+PS(0.7)	0.531	0.067	0.042	0.042	-0.036	0.520
FedCvg-Ratio(0.5)+LD+PS(0.5)	0.507	0.071	0.038	0.039	-0.034	0.490
FedCvg-Ratio(0.9)+LD+PS(0.5)	0.507	0.071	0.037	0.040	-0.035	0.490
FedCvg-Ratio(0.5)+LD+PS(0.7)	0.508	0.070	0.037	0.043	-0.038	0.490
FedCvg-Ratio(0.9)+LD+PS(0.7)	0.508	0.073	0.038	0.037	-0.034	0.491

Table D.2: Compas dataset: Performance at coverage=474, partition=same_size (best EOD across learning rates, averaged over 5 seeds)

Method	Acc	EOD	SPD	AOD	AccDiff	Prec
FairFed(EOD)	0.519	0.035	0.040	0.031	-0.044	0.478
FairFed(SPD)	0.519	0.037	0.041	0.032	-0.046	0.476
FairFed(AccDiff)	0.517	0.044	0.039	0.029	-0.034	0.484
FairFed(EOD)+LD	0.517	0.034	**0.038**	0.029	-0.042	0.472
FairFed(SPD)+LD	0.519	0.037	0.040	0.031	-0.045	0.477
FairFed(AccDiff)+LD	0.517	0.049	0.042	0.032	-0.033	0.485
FairFed(EOD)+PS(0.5)	0.517	0.039	0.040	0.029	-0.035	0.483
FairFed(SPD)+PS(0.5)	0.517	0.036	**0.037**	**0.028**	-0.037	0.481

Continued on next page

Table D.2 – continued from previous page

Method	Acc	EOD	SPD	AOD	AccDiff	Prec
FairFed(AccDiff)+PS(0.5)	0.518	0.043	0.039	0.029	-0.032	0.486
FairFed(EOD)+PS(0.7)	**0.560**	0.032	0.052	0.034	-0.069	**0.524**
FairFed(SPD)+PS(0.7)	0.518	0.036	0.040	0.030	-0.037	0.481
FairFed(AccDiff)+PS(0.7)	0.518	0.043	0.040	0.029	-0.033	0.487
FairFed(EOD)+LD+PS(0.5)	0.517	0.037	0.039	0.028	-0.038	0.483
FairFed(SPD)+LD+PS(0.5)	0.519	0.036	0.039	0.028	-0.040	0.481
FairFed(AccDiff)+LD+PS(0.5)	0.518	0.039	0.039	0.029	-0.032	0.485
FairFed(EOD)+LD+PS(0.7)	0.518	0.038	0.041	0.030	-0.037	0.482
FairFed(SPD)+LD+PS(0.7)	0.517	0.037	0.038	**0.028**	-0.039	0.479
FairFed(AccDiff)+LD+PS(0.7)	0.518	0.039	0.038	0.028	-0.032	0.486
FedAvg	0.514	0.045	0.039	0.029	-0.029	0.482
FedAvg+LD	0.515	0.045	0.039	0.030	-0.029	0.483
FedAvg+PS(0.5)	0.515	0.044	0.039	0.029	-0.028	0.484
FedAvg+PS(0.7)	0.515	0.044	**0.037**	**0.028**	-0.027	0.484
FedAvg+LD+PS(0.5)	0.516	0.044	**0.038**	0.028	-0.027	0.485
FedAvg+LD+PS(0.7)	0.514	0.044	**0.037**	**0.028**	-0.027	0.483
FedCvg(0.0001)	0.528	0.038	0.069	0.050	-0.064	0.512
FedCvg(0.001)	0.526	0.035	0.067	0.048	-0.060	0.509
FedCvg(0.0001)+LD	0.527	0.037	0.069	0.050	-0.065	0.512
FedCvg(0.001)+LD	0.527	0.035	0.067	0.048	-0.060	0.509
FedCvg(0.0001)+PS(0.5)	0.528	0.041	0.070	0.052	-0.065	0.510
FedCvg(0.001)+PS(0.5)	0.527	0.035	0.066	0.048	-0.061	0.509
FedCvg(0.0001)+PS(0.7)	0.524	0.033	0.065	0.048	-0.068	0.511
FedCvg(0.001)+PS(0.7)	0.527	0.035	0.066	0.048	-0.060	0.510
FedCvg(0.0001)+LD+PS(0.5)	0.523	0.043	0.063	0.048	-0.063	0.510
FedCvg(0.001)+LD+PS(0.5)	0.527	0.035	0.066	0.048	-0.060	0.509
FedCvg(0.0001)+LD+PS(0.7)	0.527	0.038	0.067	0.049	-0.062	0.511
FedCvg(0.001)+LD+PS(0.7)	0.526	0.035	0.066	0.048	-0.062	0.509
FedCvg-Ratio(0.5)	0.524	0.031	0.063	0.048	**-0.076**	0.517
FedCvg-Ratio(0.9)	0.524	0.031	0.063	0.048	**-0.076**	0.517
FedCvg-Ratio(0.5)+LD	0.525	**0.030**	0.063	0.047	-0.075	0.518
FedCvg-Ratio(0.9)+LD	0.525	**0.030**	0.063	0.047	-0.075	0.518
FedCvg-Ratio(0.5)+PS(0.5)	0.531	0.034	0.072	0.054	-0.073	0.518
FedCvg-Ratio(0.9)+PS(0.5)	0.529	0.036	0.072	0.055	-0.071	0.513
FedCvg-Ratio(0.5)+PS(0.7)	0.530	0.036	0.074	0.056	-0.071	0.514
FedCvg-Ratio(0.9)+PS(0.7)	0.531	0.034	0.074	0.055	-0.073	0.517
FedCvg-Ratio(0.5)+LD+PS(0.5)	0.528	0.038	0.074	0.055	-0.072	0.511
FedCvg-Ratio(0.9)+LD+PS(0.5)	0.531	0.039	0.076	0.058	-0.072	0.516
FedCvg-Ratio(0.5)+LD+PS(0.7)	0.525	0.034	0.066	0.051	-0.074	0.519
FedCvg-Ratio(0.9)+LD+PS(0.7)	0.524	0.035	0.067	0.051	-0.074	0.516

Table D.3: Compas dataset: Performance at coverage=610, partition=diff_size (best EOD across learning rates, averaged over 5 seeds)

Method	Acc	EOD	SPD	AOD	AccDiff	Prec
FairFed(EOD)	**0.532**	0.059	0.050	0.037	-0.035	**0.535**
FairFed(SPD)	0.530	0.059	0.050	0.039	-0.039	0.531
FairFed(AccDiff)	0.528	**0.050**	0.047	0.034	-0.040	0.525
FairFed(EOD)+LD	**0.532**	0.059	0.051	0.039	-0.037	**0.535**
FairFed(SPD)+LD	0.530	0.056	0.052	0.039	-0.040	0.530
FairFed(AccDiff)+LD	0.528	0.050	0.048	0.034	-0.039	0.524
FairFed(EOD)+PS(0.5)	0.530	0.058	0.048	0.036	-0.039	0.531
FairFed(SPD)+PS(0.5)	0.529	0.051	0.047	0.034	-0.039	0.524
FairFed(AccDiff)+PS(0.5)	0.529	0.051	0.048	0.035	-0.039	0.527
FairFed(EOD)+PS(0.7)	0.531	0.058	0.050	0.037	-0.037	0.532
FairFed(SPD)+PS(0.7)	0.530	0.060	0.051	0.040	-0.036	0.532
FairFed(AccDiff)+PS(0.7)	0.529	0.056	0.051	0.038	-0.039	0.525
FairFed(EOD)+LD+PS(0.5)	0.530	0.055	0.052	0.038	-0.038	0.526
FairFed(SPD)+LD+PS(0.5)	0.530	0.052	0.046	0.033	-0.037	0.529
FairFed(AccDiff)+LD+PS(0.5)	0.529	**0.050**	0.045	0.033	-0.035	0.521
FairFed(EOD)+LD+PS(0.7)	0.530	0.057	0.051	0.039	-0.038	0.530
FairFed(SPD)+LD+PS(0.7)	0.530	0.058	0.053	0.039	-0.036	0.532
FairFed(AccDiff)+LD+PS(0.7)	0.527	0.050	0.049	0.036	-0.039	0.521
FedAvg	0.531	0.061	0.051	0.039	-0.036	0.532
FedAvg+LD	0.530	0.058	0.051	0.038	-0.038	0.529
FedAvg+PS(0.5)	0.530	0.052	0.046	0.033	-0.037	0.526
FedAvg+PS(0.7)	0.530	0.054	0.049	0.036	-0.037	0.527
FedAvg+LD+PS(0.5)	0.530	**0.050**	0.045	**0.031**	-0.038	0.527
FedAvg+LD+PS(0.7)	0.530	0.055	0.050	0.037	-0.039	0.527
FedCvg(0.0001)	0.521	0.066	0.033	0.037	-0.038	0.517
FedCvg(0.001)	0.523	0.066	0.036	0.040	-0.042	0.520
FedCvg(0.0001)+LD	0.521	0.062	**0.031**	0.036	-0.039	0.518
FedCvg(0.001)+LD	0.523	0.068	0.035	0.040	-0.040	0.520
FedCvg(0.0001)+PS(0.5)	0.522	0.065	0.033	0.037	-0.038	0.519
FedCvg(0.001)+PS(0.5)	0.521	0.065	0.033	0.037	-0.041	0.520
FedCvg(0.0001)+PS(0.7)	0.521	0.063	0.032	0.036	-0.039	0.518
FedCvg(0.001)+PS(0.7)	0.522	0.069	0.035	0.040	-0.039	0.520
FedCvg(0.0001)+LD+PS(0.5)	0.521	0.062	0.032	0.036	-0.041	0.517
FedCvg(0.001)+LD+PS(0.5)	0.522	0.066	**0.032**	0.037	-0.038	0.521
FedCvg(0.0001)+LD+PS(0.7)	0.522	0.063	0.032	0.036	-0.039	0.521
FedCvg(0.001)+LD+PS(0.7)	0.522	0.067	0.034	0.038	-0.040	0.518
FedCvg-Ratio(0.5)	0.521	0.070	0.037	0.042	-0.038	0.518
FedCvg-Ratio(0.9)	0.521	0.070	0.037	0.042	-0.038	0.518
FedCvg-Ratio(0.5)+LD	0.521	0.064	0.034	0.039	-0.039	0.518

Continued on next page

Table D.3 – continued from previous page

Method	Acc	EOD	SPD	AOD	AccDiff	Prec
FedCvg-Ratio(0.9)+LD	0.521	0.064	0.034	0.039	-0.039	0.518
FedCvg-Ratio(0.5)+PS(0.5)	0.520	0.064	0.033	0.038	-0.041	0.515
FedCvg-Ratio(0.9)+PS(0.5)	0.520	0.068	0.034	0.040	-0.040	0.516
FedCvg-Ratio(0.5)+PS(0.7)	0.521	0.068	0.034	0.039	-0.040	0.518
FedCvg-Ratio(0.9)+PS(0.7)	0.520	0.068	0.035	0.039	-0.040	0.516
FedCvg-Ratio(0.5)+LD+PS(0.5)	0.520	0.060	**0.031**	0.035	-0.042	0.515
FedCvg-Ratio(0.9)+LD+PS(0.5)	0.518	0.063	0.033	0.038	**-0.043**	0.512
FedCvg-Ratio(0.5)+LD+PS(0.7)	0.521	0.065	0.034	0.038	-0.041	0.517
FedCvg-Ratio(0.9)+LD+PS(0.7)	0.520	0.064	0.033	0.037	-0.040	0.518

Table D.4: Compas dataset: Performance at coverage=610, partition=same_size (best EOD across learning rates, averaged over 5 seeds)

Method	Acc	EOD	SPD	AOD	AccDiff	Prec
FairFed(EOD)	**0.549**	0.043	0.031	0.026	-0.043	0.554
FairFed(SPD)	0.542	0.037	0.034	0.021	-0.014	0.517
FairFed(AccDiff)	0.543	0.043	0.038	0.023	-0.036	**0.558**
FairFed(EOD)+LD	**0.549**	0.045	0.033	0.028	-0.044	0.555
FairFed(SPD)+LD	0.543	**0.033**	0.033	0.017	-0.013	0.519
FairFed(AccDiff)+LD	0.531	0.043	0.032	0.020	-0.029	0.525
FairFed(EOD)+PS(0.5)	0.532	0.037	0.024	0.015	-0.027	0.525
FairFed(SPD)+PS(0.5)	0.537	0.040	0.036	0.022	-0.025	0.521
FairFed(AccDiff)+PS(0.5)	0.532	0.041	0.030	0.018	-0.026	0.530
FairFed(EOD)+PS(0.7)	0.522	0.037	0.017	0.013	-0.027	0.488
FairFed(SPD)+PS(0.7)	0.537	0.045	0.034	0.022	-0.030	0.519
FairFed(AccDiff)+PS(0.7)	0.533	0.043	0.030	0.018	-0.028	0.528
FairFed(EOD)+LD+PS(0.5)	0.530	**0.034**	0.020	**0.013**	-0.027	0.520
FairFed(SPD)+LD+PS(0.5)	0.537	0.044	0.033	0.020	-0.026	0.518
FairFed(AccDiff)+LD+PS(0.5)	0.529	0.044	0.027	0.017	-0.025	0.520
FairFed(EOD)+LD+PS(0.7)	0.521	0.036	**0.017**	**0.013**	-0.027	0.484
FairFed(SPD)+LD+PS(0.7)	0.521	0.045	0.025	0.016	-0.017	0.488
FairFed(AccDiff)+LD+PS(0.7)	0.531	0.041	0.029	0.017	-0.028	0.525
FedAvg	0.521	0.039	0.021	0.016	-0.014	0.489
FedAvg+LD	0.520	0.039	0.021	0.015	-0.014	0.489
FedAvg+PS(0.5)	0.520	0.037	0.023	0.017	-0.014	0.488
FedAvg+PS(0.7)	0.520	0.039	0.021	0.015	-0.014	0.489
FedAvg+LD+PS(0.5)	0.520	0.038	0.021	0.016	-0.014	0.489
FedAvg+LD+PS(0.7)	0.521	0.039	0.021	0.015	-0.013	0.489
FedCvg(0.0001)	0.531	0.049	0.058	0.041	-0.045	0.530
FedCvg(0.001)	0.516	0.049	0.033	0.026	-0.021	0.509

Continued on next page

Table D.4 – continued from previous page

Method	Acc	EOD	SPD	AOD	AccDiff	Prec
FedCvg(0.0001)+LD	0.515	0.049	0.033	0.025	-0.027	0.508
FedCvg(0.001)+LD	0.516	0.049	0.033	0.025	-0.021	0.508
FedCvg(0.0001)+PS(0.5)	0.531	0.049	0.059	0.042	-0.046	0.530
FedCvg(0.001)+PS(0.5)	0.531	0.049	0.060	0.043	-0.041	0.526
FedCvg(0.0001)+PS(0.7)	0.539	0.048	0.047	0.032	-0.027	0.531
FedCvg(0.001)+PS(0.7)	0.515	0.048	0.033	0.024	-0.022	0.507
FedCvg(0.0001)+LD+PS(0.5)	0.530	0.048	0.059	0.042	-0.045	0.527
FedCvg(0.001)+LD+PS(0.5)	0.516	0.049	0.033	0.026	-0.024	0.508
FedCvg(0.0001)+LD+PS(0.7)	0.530	0.050	0.061	0.043	-0.047	0.532
FedCvg(0.001)+LD+PS(0.7)	0.516	0.049	0.031	0.024	-0.022	0.509
FedCvg-Ratio(0.5)	0.532	0.037	0.060	0.039	**-0.056**	0.532
FedCvg-Ratio(0.9)	0.532	0.037	0.060	0.039	**-0.056**	0.532
FedCvg-Ratio(0.5)+LD	0.532	0.038	0.057	0.037	-0.055	0.532
FedCvg-Ratio(0.9)+LD	0.532	0.038	0.057	0.037	-0.055	0.532
FedCvg-Ratio(0.5)+PS(0.5)	0.532	0.045	0.060	0.040	-0.051	0.532
FedCvg-Ratio(0.9)+PS(0.5)	0.532	0.040	0.058	0.038	-0.055	0.534
FedCvg-Ratio(0.5)+PS(0.7)	0.532	0.039	0.057	0.037	-0.055	0.535
FedCvg-Ratio(0.9)+PS(0.7)	0.531	0.044	0.059	0.039	-0.049	0.529
FedCvg-Ratio(0.5)+LD+PS(0.5)	0.531	0.044	0.059	0.040	-0.051	0.532
FedCvg-Ratio(0.9)+LD+PS(0.5)	0.532	0.040	0.058	0.038	-0.055	0.535
FedCvg-Ratio(0.5)+LD+PS(0.7)	0.529	0.046	0.053	0.036	-0.051	0.544
FedCvg-Ratio(0.9)+LD+PS(0.7)	0.533	0.042	0.062	0.042	-0.053	0.537

Table D.5: Compas dataset: Performance at coverage=1067, partition=diff_size (best EOD across learning rates, averaged over 5 seeds)

Method	Acc	EOD	SPD	AOD	AccDiff	Prec
FairFed(EOD)	**0.571**	0.064	0.052	0.040	-0.020	0.576
FairFed(SPD)	**0.571**	0.067	0.057	0.042	-0.018	0.581
FairFed(AccDiff)	0.569	0.053	0.055	0.035	-0.026	0.584
FairFed(EOD)+LD	0.569	0.061	0.054	0.036	-0.017	0.573
FairFed(SPD)+LD	0.570	0.065	0.057	0.039	-0.016	0.581
FairFed(AccDiff)+LD	0.568	**0.051**	0.053	0.034	**-0.028**	0.582
FairFed(EOD)+PS(0.5)	0.569	0.055	0.056	0.037	**-0.028**	0.581
FairFed(SPD)+PS(0.5)	0.567	0.068	0.060	0.044	-0.023	0.579
FairFed(AccDiff)+PS(0.5)	0.565	0.053	0.060	0.040	-0.026	0.578
FairFed(EOD)+PS(0.7)	0.563	0.061	0.058	0.041	-0.021	0.564
FairFed(SPD)+PS(0.7)	0.570	0.064	0.060	0.041	-0.017	0.577
FairFed(AccDiff)+PS(0.7)	0.566	0.053	0.060	0.040	**-0.028**	0.582
FairFed(EOD)+LD+PS(0.5)	0.567	0.061	0.057	0.039	-0.018	0.571
FairFed(SPD)+LD+PS(0.5)	0.564	0.072	0.061	0.046	-0.018	0.571

Continued on next page

Table D.5 – continued from previous page

Method	Acc	EOD	SPD	AOD	AccDiff	Prec
FairFed(AccDiff)+LD+PS(0.5)	0.566	0.054	0.056	0.037	-0.025	0.581
FairFed(EOD)+LD+PS(0.7)	0.569	0.059	0.056	0.038	-0.025	**0.586**
FairFed(SPD)+LD+PS(0.7)	0.566	0.070	0.059	0.044	-0.018	0.570
FairFed(AccDiff)+LD+PS(0.7)	0.567	0.058	0.056	0.039	-0.025	0.579
FedAvg	0.569	0.063	0.061	0.042	-0.019	0.581
FedAvg+LD	0.569	0.064	0.059	0.041	-0.017	0.580
FedAvg+PS(0.5)	0.566	0.063	0.059	0.043	-0.021	0.574
FedAvg+PS(0.7)	0.567	0.061	0.058	0.043	-0.024	0.576
FedAvg+LD+PS(0.5)	0.567	0.064	0.057	0.041	-0.022	0.578
FedAvg+LD+PS(0.7)	0.567	0.063	0.058	0.042	-0.021	0.577
FedCvg(0.0001)	0.561	0.064	0.052	0.039	-0.012	0.538
FedCvg(0.001)	0.564	0.068	0.054	0.042	-0.010	0.539
FedCvg(0.0001)+LD	0.560	0.063	0.050	0.038	-0.012	0.536
FedCvg(0.001)+LD	0.563	0.069	0.051	0.040	-0.007	0.547
FedCvg(0.0001)+PS(0.5)	0.558	0.064	0.049	0.036	-0.010	0.532
FedCvg(0.001)+PS(0.5)	0.559	0.060	0.050	0.038	-0.012	0.532
FedCvg(0.0001)+PS(0.7)	0.558	0.061	0.046	0.034	-0.011	0.534
FedCvg(0.001)+PS(0.7)	0.561	0.061	0.051	0.038	-0.013	0.536
FedCvg(0.0001)+LD+PS(0.5)	0.556	0.062	0.046	0.034	-0.008	0.529
FedCvg(0.001)+LD+PS(0.5)	0.559	0.061	0.051	0.038	-0.011	0.531
FedCvg(0.0001)+LD+PS(0.7)	0.560	0.061	0.046	0.033	-0.009	0.535
FedCvg(0.001)+LD+PS(0.7)	0.561	0.060	0.049	0.036	-0.010	0.535
FedCvg-Ratio(0.5)	0.553	0.061	0.043	0.031	-0.017	0.529
FedCvg-Ratio(0.9)	0.553	0.061	0.043	0.031	-0.017	0.529
FedCvg-Ratio(0.5)+LD	0.552	0.063	0.039	**0.027**	-0.014	0.528
FedCvg-Ratio(0.9)+LD	0.552	0.063	0.039	**0.027**	-0.014	0.528
FedCvg-Ratio(0.5)+PS(0.5)	0.552	0.064	0.040	0.029	-0.015	0.528
FedCvg-Ratio(0.9)+PS(0.5)	0.553	0.063	**0.039**	**0.027**	-0.012	0.527
FedCvg-Ratio(0.5)+PS(0.7)	0.553	0.063	0.040	**0.028**	-0.016	0.529
FedCvg-Ratio(0.9)+PS(0.7)	0.553	0.061	0.041	0.029	-0.017	0.529
FedCvg-Ratio(0.5)+LD+PS(0.5)	0.553	0.070	0.040	0.029	-0.011	0.527
FedCvg-Ratio(0.9)+LD+PS(0.5)	0.551	0.069	0.042	0.031	-0.010	0.526
FedCvg-Ratio(0.5)+LD+PS(0.7)	0.552	0.067	**0.039**	**0.028**	-0.012	0.527
FedCvg-Ratio(0.9)+LD+PS(0.7)	0.552	0.069	**0.039**	0.029	-0.011	0.527

Table D.6: Compas dataset: Performance at coverage=1067, partition=same_size (best EOD across learning rates, averaged over 5 seeds)

Method	Acc	EOD	SPD	AOD	AccDiff	Prec
FairFed(EOD)	0.572	0.066	0.061	0.050	-0.031	0.577
FairFed(SPD)	**0.588**	0.073	0.072	0.057	-0.034	**0.607**
FairFed(AccDiff)	0.521	0.073	**0.036**	**0.041**	-0.011	0.497
FairFed(EOD)+LD	0.572	0.062	0.056	0.045	-0.030	0.582
FairFed(SPD)+LD	**0.588**	0.070	0.071	0.056	**-0.035**	**0.607**
FairFed(AccDiff)+LD	0.572	0.069	0.053	0.045	-0.023	0.582
FairFed(EOD)+PS(0.5)	0.535	0.072	0.050	0.049	-0.026	0.526
FairFed(SPD)+PS(0.5)	0.537	0.074	0.050	0.049	-0.023	0.532
FairFed(AccDiff)+PS(0.5)	0.574	0.073	0.059	0.050	-0.024	0.579
FairFed(EOD)+PS(0.7)	0.571	0.073	0.060	0.051	-0.025	0.578
FairFed(SPD)+PS(0.7)	0.537	0.073	0.049	0.049	-0.024	0.529
FairFed(AccDiff)+PS(0.7)	0.522	0.074	0.037	0.042	-0.014	0.498
FairFed(EOD)+LD+PS(0.5)	0.522	0.077	0.041	0.048	-0.018	0.496
FairFed(SPD)+LD+PS(0.5)	0.536	0.072	0.049	0.048	-0.024	0.527
FairFed(AccDiff)+LD+PS(0.5)	0.571	0.071	0.058	0.049	-0.025	0.580
FairFed(EOD)+LD+PS(0.7)	0.524	0.070	0.040	0.045	-0.022	0.496
FairFed(SPD)+LD+PS(0.7)	0.538	0.079	0.052	0.052	-0.021	0.536
FairFed(AccDiff)+LD+PS(0.7)	0.571	0.070	0.057	0.048	-0.025	0.578
FedAvg	0.537	0.076	0.053	0.053	-0.024	0.526
FedAvg+LD	0.536	0.074	0.052	0.053	-0.025	0.525
FedAvg+PS(0.5)	0.536	0.074	0.049	0.050	-0.022	0.525
FedAvg+PS(0.7)	0.537	0.076	0.053	0.053	-0.024	0.526
FedAvg+LD+PS(0.5)	0.535	0.071	0.050	0.049	-0.024	0.522
FedAvg+LD+PS(0.7)	0.535	0.074	0.053	0.053	-0.025	0.519
FedCvg(0.0001)	0.535	0.059	0.063	0.052	-0.013	0.549
FedCvg(0.001)	0.532	0.052	0.063	0.051	-0.005	0.537
FedCvg(0.0001)+LD	0.534	0.060	0.064	0.054	-0.013	0.547
FedCvg(0.001)+LD	0.532	0.052	0.063	0.051	-0.005	0.537
FedCvg(0.0001)+PS(0.5)	0.535	0.052	0.062	0.050	-0.012	0.542
FedCvg(0.001)+PS(0.5)	0.533	**0.047**	0.063	0.051	-0.013	0.546
FedCvg(0.0001)+PS(0.7)	0.534	0.059	0.064	0.053	-0.014	0.547
FedCvg(0.001)+PS(0.7)	0.533	0.051	0.063	0.051	-0.007	0.541
FedCvg(0.0001)+LD+PS(0.5)	0.536	0.053	0.063	0.051	-0.014	0.541
FedCvg(0.001)+LD+PS(0.5)	0.534	0.052	0.062	0.051	-0.007	0.543
FedCvg(0.0001)+LD+PS(0.7)	0.536	0.054	0.063	0.051	-0.012	0.543
FedCvg(0.001)+LD+PS(0.7)	0.533	0.052	0.064	0.053	-0.007	0.541
FedCvg-Ratio(0.5)	0.537	0.056	0.057	0.046	-0.022	0.549
FedCvg-Ratio(0.9)	0.537	0.056	0.057	0.046	-0.022	0.549
FedCvg-Ratio(0.5)+LD	0.534	0.053	0.054	0.044	-0.024	0.541

Continued on next page

Table D.6 – continued from previous page

Method	Acc	EOD	SPD	AOD	AccDiff	Prec
FedCvg-Ratio(0.9)+LD	0.534	0.053	0.054	0.044	-0.024	0.541
FedCvg-Ratio(0.5)+PS(0.5)	0.535	0.055	0.059	0.048	-0.019	0.545
FedCvg-Ratio(0.9)+PS(0.5)	0.534	0.056	0.058	0.048	-0.018	0.552
FedCvg-Ratio(0.5)+PS(0.7)	0.534	0.058	0.059	0.048	-0.017	0.544
FedCvg-Ratio(0.9)+PS(0.7)	0.534	0.059	0.061	0.051	-0.018	0.540
FedCvg-Ratio(0.5)+LD+PS(0.5)	0.534	0.054	0.058	0.048	-0.022	0.541
FedCvg-Ratio(0.9)+LD+PS(0.5)	0.532	0.050	0.060	0.049	-0.017	0.541
FedCvg-Ratio(0.5)+LD+PS(0.7)	0.534	0.055	0.059	0.049	-0.021	0.544
FedCvg-Ratio(0.9)+LD+PS(0.7)	0.533	0.054	0.057	0.047	-0.020	0.541