MSc Artificial Intelligence
Master Thesis

Bridging Fairness and Privacy: An Experimental Analysis of Bias Assessment in Federated Learning

by
Jelke Matthijsse
12653500

June 29, 2024

36 EC
January - June 2024

Supervisors:
Dr. G. Sileno (UvA)
I. Barberá (Rhite)

Examiner:
Dr. G. Sileno

Second reader:
D. Chabal MSc

Universiteit van Amsterdam

“The question really isn’t ‘Is AI as smart as us?’
it’s more ‘Are we as dumb as AI?’”

– Hank Green

Acknowledgement

First and foremost, I extend my gratitude to my thesis supervisor, Dr. Giovanni Sileno for his guidance and encouragement throughout this project. His invaluable insights and feedback during our weekly meetings were instrumental in shaping the direction of this thesis. I would also like to thank Daphnée Chabal, for taking the time to be the second examinator of this thesis.

Furthermore, I am deeply indebted to the team at Rhite for their unwavering support and the resources they provided, which formed the foundation of this research. In particular, I want to extend my gratitude to Isabel Barberá and Shieltaa Dielbandhoesing for their consistent feedback and assistance over the past few months.

A special thanks goes to my fellow students for fostering a stimulating working environment, which greatly enriched this experience.

On a personal note, I want to thank my brother Jip for proofreading this thesis and providing answers to all my research-related questions.

Abstract

The increasing integration of Artificial Intelligence (AI) in crucial decision-making areas, such as healthcare and recruitment, underscores the importance of developing AI models that ensure fairness across various demographic groups and safeguard the confidentiality of personal data. Federated learning (FL) has emerged as a privacy-preserving, decentralized method that trains a global model by aggregating locally trained models, thereby maintaining local data privacy. However, FL brings new challenges in detecting bias due to limited access to training data. Previous research has proposed an aggregated local bias assessment method for FL that aggregates local bias scores using the same aggregation algorithm that is used for model aggregation. However, there lacks a theoretical basis and comprehensive experimental evaluation of this method. This thesis provides an experimental analysis of bias assessment within federated learning, thereby focusing on two main objectives: (1) evaluating the accuracy of the privacy-preserving aggregated local bias assessment technique, and (2) comparing bias that arises in an FL model to bias that arises in a centrally trained model. The first research objective is addressed by comparing the bias in an FL model measured by the aggregated local bias assessment with the bias measured by a global bias assessment that disregards privacy. Results showed that the aggregated local assessment detected more bias than the global assessment in sensitive attributes with unequally distributed classes. The second objective is addressed by comparing bias from a centrally trained model with the globally-measured bias from an FL model. These experiments revealed discrepancies between the central pipeline and the FL pipeline. However, these discrepancies did not consistently indicate more bias for the FL pipeline. Additionally, the influence of client heterogeneity on both research objectives was researched by introducing experimental client partitions that entail different types and amounts of data heterogeneity. Findings revealed that neither the FL model with aggregated local assessment nor a global bias assessment showed consistent differences in bias scores due to heterogeneity, suggesting that heterogeneity does not affect the accuracy of the aggregated local bias assessment method or the bias that arises during federated learning.

Contents

Chapter 1

Introduction

As humans, we make a lot of choices every day. These choices can vary from being very simple, like choosing what movie to watch, to very complex and impactful, like who to hire or what sentence someone on trial should receive. In the last few years, *Artificial Intelligence* (AI) is more frequently adopted to assist in decision-making tasks that have an impact on people's lifes (Russell and Norvig, 2016; Smith and Eckroth, 2017). Machine learning is able to quickly learn complex patterns from data, thereby allowing for efficient automation and guidance of tasks and decision-making processes in areas such as credit scoring (Dastile et al., 2020), recruitment (Faliagka et al., 2012) and healthcare (Jayatilake et al., 2021). However, as AI increasingly influences critical decisions that affect humans, the need for responsible, safe, and transparent AI becomes more relevant to protect the rights and interests of individuals.

Despite the 'rational' aim of AI, using machine learning for personal decisions introduces the risk of disproportionate treatment between individuals or different demographic groups. For instance, data samples of minority groups are often underrepresented because they are deemed less crucial to improve the accuracy of a model compared to datapoints from majority groups. Consequently, the issue of *unfairness* within AI models has gained significant interest by research and industry (Mehrabi et al., 2021b). In the context of decision-making, fairness is often referred to as an absence of any prejudice or favoritism towards (groups of) individuals (Mehrabi et al., 2021b) and is typically measured by the presence of *negative bias*. Unfairness in AI systems can be caused by training on biased data (e.g., imbalanced or unrepresentative data) or by bias that arises from the learning algorithms themselves.

The emergence of AI models that learn from big data also poses challenges related to privacy protection (Khanan et al., 2019). Training machine learning on personal data - as is often the case when AI is used for human-centric tasks - has raised concerns about the leakage of individuals personal data through the training process (Carmody et al., 2021). These concerns have led to a new interest in the development of AI models under privacy constraints. Current research tries to ensure privacy in AI models by exploiting methods like differential privacy (Zhao and Chen, 2022), federated learning (Mammen, 2021) and cryptography (Zapechnikov, 2022).

A holistic view of a safe and responsible AI system, one that minimizes risks and harms to its users, should address not only bias and discrimination but also the safe and discreet use of users' personal data (Jobin et al., 2019). Most work on responsible AI focuses on either *fair learning*, which ensures equal treatment and outcomes across different user groups, or on *private learning*, which protects the confidentiality of users' data during the learning process. Recently, work on the combined field of *private fair learning* has sparked interest within research (e.g. Ferraguig et al., 2021; Fioretto et al., 2022; Chang and Shokri, 2023; Shaham et al., 2023).

1.1 Problem Description

One commonly-used private learning method is *federated learning* (FL). This method ensures privacy through decentralized learning by employing different *local* models that are trained with local datasets. These locally trained models are then aggregated to form a *global* model, which can be done with different FL aggregation algorithms that can differ in *what* (e.g. gradient updates or model weights), *when* (e.g. synchronous or asynchronous across clients) and *how* (e.g. weighted average, etc.) model updates are aggregated. Federated learning ensures that local data will not leave its original location, and thereby aims to keep local data private. Federated learning can, for instance, be applied in the medical field to train a model that helps with medical decision-making, without compromising sensitive patient data that is often spread across different hospitals. Despite the potential of this privacy-by-design approach, recent work has shown that federated learning can still be vulnerable to privacy attacks, due to the model update sharing among clients (Truong et al., 2021).

Nonetheless, the privacy-focused and decentralized nature of federated learning makes this method a promising technique for practical applications in industry. Data often originates from different isolated sources that are disconnected from each other. This configuration, known as *data islands* (or *data silos*) makes the centralized collection of data a complex, time-consuming and expensive task (Yu et al., 2022). Federated learning addresses this issue and is therefore already being utilized in real-world settings by companies such as Google and IBM, as well as institutions in healthcare and finance (Li et al., 2020).

However, regardless of the advantages of federated learning, it has been shown that this method can give rise to bias (Kairouz et al., 2021; Chang and Shokri, 2023). A holistic view on safe and responsible AI should consider both a safe use of personal data and possible harms that arise from biased and unfair treatment (Jobin et al., 2019). Most existing methods for detecting and mitigating bias are designed for centralized learning settings. These methods often require access to the complete dataset, which is not possible in federated learning. Addressing fairness with the extra privacy dimension realized by federated learning, still remains a less-researched and open question.

Although the topic of federated fair learning has been previously researched, the extent of this research is limited. Many studies directly assume that bias arises during federated learning. These studies propose privacy-preserving bias mitigation methods that address fairness without analyzing the actual bias that is caused by federated learning further (Zhang et al., 2020; Abay et al., 2020; Zeng et al., 2021; Ezzeldin et al., 2023). However, before mitigating bias, it is crucial to understand how and where bias arises with the help of bias assessment techniques. Additionally, bias assessment in federated learning requires methods that can detect bias without compromising the local data privacy during the process. This means that before actually assessing the bias within the federated learning framework, it is important to define and evaluate techniques that can do this without compromising local privacy, a problem that the current literature has failed to address.

Some research on bias mitigation in federated learning use an aggregation of local bias scores to obtain a bias score for the global FL model (Zhang et al., 2020; Ezzeldin et al., 2023). This is typically done by following the same aggregation algorithm (e.g., weighted average) by which the model updates are aggregated. However, a thorough review of the literature reveals a lack of both theoretical analysis and of comprehensive experimental evaluation for this approach. It is therefore not proven that the aggregation of local bias scores will always result in a correct global bias score that is representative of the bias in the global FL model with respect to all data. Moreover, using the same aggregation algorithm for bias assessment as is used for model aggregation seems intuitively problematic, as this algorithm is the one needing assessment for bias in the first place.

As a consequence, the focus of this research will be two-fold. First, we will assess the correctness of the currently used *aggregated local bias assessment* approach in federated learning. Secondly, this research will assess bias that can be introduced by the federated learning framework (e.g. bias that is caused by the federated aggregation algorithm). For simplicity, this research considers only *group fairness*, the most studied type of algorithmic fairness at the moment, and thus focuses on bias that arises between different demographic groups. Therefore, the first research question will be: *How accurate is the aggregation bias assessment method in detecting bias that arises in a federated learning framework?* This question will be answered using the following sub-questions.

- How is bias currently detected in a federated learning framework using a local aggregation method?
- How does the local aggregation bias detection method compare to bias detection methods that do not preserve privacy?
- How does data heterogeneity affect the bias measured by the local aggregation method?

After assessing the functionality of this bias assessment approach, the second part of this research focuses on the bias that arises during the federated learning process. This leads to the second research question: *What bias arises during the federated learning process compared to a centralized learning process?* This question will be answered using the following sub-questions.

- How can bias detection methods be used to detect bias within a federated learning pipeline?
- How does data heterogeneity influence the bias that arises during a federated learning pipeline?

1.2 Thesis Overview

Chapter 2 will provide an overview of the main concepts of this thesis and how the proposed research can be positioned in already existing work on the topic. Chapter 3 discusses the proposed methodology for this research, and Chapter 4 discusses the experimental setup in which the research will be conducted. The results will then be presented in Chapter 5 and further analyzed in Chapter 6. Finally, a conclusive evaluation will be given in Chapter 7, alongside directions for future work.

Chapter 2

Background and Related Work

This chapter provides an overview of the foundational concepts and existing research that contextualize the field of fairness in federated learning. This chapter first discusses the concept of fairness in Section 2.1, thereby providing the definitions of fairness and bias used in this study. Section 2.2 explores the concept of privacy and common privacy-preserving learning methods, including federated learning. In Section 2.3 we review the current literature on fairness in federated learning. Finally, Section 2.4 identifies the research gaps within the field and how these will be addressed by this thesis.

2.1 Fairness

The concept of fairness has a wide variety of meanings in literature, media and society (Verma and Rubin, 2018). In the context of AI, it is crucial to establish a clear definition of fairness that can be applied to computational AI systems. This section will provide coherent definitions of fairness and bias that will be used in the remainder of this thesis.

2.1.1 Definition of Fairness

Fairness is a fluid concept that can change over time and that can have different meanings within different contexts and societies. What is considered fair in one era or culture might not be regarded as such in another, which highlights the importance of understanding fairness as a dynamic and context-dependent notion. Within the context of decision-making, fairness is often referred to as the *absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics* (Mehrabi et al., 2021b). According to this definition, an unfair algorithm is an algorithm whose decisions are biased towards a particular group of people. In the field of AI, characteristics that might lead to individual discrimination are often referred to as *sensitive attributes.* Examples of such sensitive attributes include age, race and gender.

2.1.2 Definition of Bias

The concept of bias is closely related to fairness, and the terms are often used interchangeably. However, there are important distinctions between them. Bias is often referred to as *a systematic tendency of an algorithm that can influence the outcome for certain individuals or groups in a way that deviates from neutrality* (Ferrara, 2023). This definition shows that the problem of bias is a technical problem, while fairness is a social concept that relates to the absence of favoritism. Bias is often seen as a cause of unfairness, although bias can be both positive and negative (Ferrara, 2023). Positive bias is the tendency to deviate towards a more favorable outcome than the true expected outcome. Negative bias is the opposite scenario in which the deviation leads to a less favorable outcome than the true expected outcome. Addressing negative biases is often the main motivation to achieve fairness (Mehrabi et al., 2021b).

2.1.3 Fairness and Bias in AI

In technical contexts, the definitions of fairness and bias as given in Section 2.1.1 and 2.1.2 are further specified depending on their type and how they are measured.

Types of Fairness

Fairness definitions can be roughly divided into two main groups: definitions that consider fairness from a *statistical* perspective and definitions that consider fairness from an *individual* perspective (Chouldechova and Roth, 2018). The statistical definitions of fairness assess the parity of a statistical measurement across different demographic groups. A demographic group is a group defined by shared characteristics (e.g., race, sex or occupation). These characteristics can be protected (e.g. *female* or *Black*) or unprotected (e.g. *male*, *White* or *teacher*). The fairness definitions that look at parity between demographic groups fall under *group fairness*, thereby indicating that different groups should be treated equally to satisfy fairness. The second group of fairness definitions focuses on individuals. *Individual fairness* states that similar predictions should be given to similar individuals. These definitions focus on constraints that are used on pairs of individuals, instead of employing a quantity that is averaged over groups. There is a third, less-common type of definitions that focuses on *subgroup fairness*. Methods that fall under this category evaluate fairness across groups defined by combinations of shared characteristics (e.g. *Black* and *female*).

Fairness Metrics

There are different ways in which these types of fairness can be measured and evaluated within algorithms. In the past decade, over twenty different notions of fairness have been proposed (Verma and Rubin, 2018). Many of these commonly-used *fairness constraints* use statistical measures to obtain a practical and objective framework that can numerically assess fairness in AI models. The fairness constraints that will be used for the bias assessment within this research are *Equalized Odds* and *Demographic Parity*, due to their common application and easy implementation. These fairness constraints are detailed below. Other fairness constraints, such as *Equal Opportunity*, *Fairness Through Awareness*, *Treatment Equality*, *Test Fairness*, *Counterfactual Fairness*, *Conditional Statistical Parity*, etc. (Mehrabi et al., 2021a), are outside the scope of this research and will not be discussed. In the definitions provided below, AI models are referred to as a *predictor* (e.g. an algorithm, or learning pipeline). These definitions consider binary classification problems.

- **Equalized Odds** A predictor satisfies equalized odds if the outcome of the predictor and a sensitive attribute are conditionally independent on the true expected outcome. This means that the probability of a person in the positive class being correctly assigned a positive outcome and the probability of a person in the negative class being incorrectly assigned a positive outcome should be equal. In other words, to satisfy equalized odds, the true positive rate and false positive rate should be equal for the individuals in the protected and unprotected groups.
- **Demographic / statistical parity** A predictor satisfies demographic parity when the probability of a positive prediction is independent of a sensitive attribute. In other words,

the likelihood of a positive outcome should be equal for all protected and unprotected groups.

Types of Bias

There are many different places in the machine learning life cycle in which biases can emerge, which leads to different types of biases. This section reiterates some of the most important sources of bias as presented in (Mehrabi et al., 2021b). First, there is *historical bias*, which are biases and sociotechnical issues that already exist in the world and can seep into the dataset during the data generation process. Then there is *representation bias*, which are biases that arise from how there is sampled from a population during the data collection process. *Measurement bias* are biases that arise from choices in how to utilize, measure and decide on particular features. Then there is *deployment bias* which arises after the employment of a model, when there is interaction with real users, and is a result of change in population, cultural values or societal knowledge. There is *evaluation bias* which arises during the model evaluation due to inappropriate and/or disproportionate benchmarks. *Aggregation bias* arises when wrong conclusions are drawn for an individual by observing an entire population. And lastly, there is *algorithmic bias*, which are biases that are added by the algorithm.

Bias Assessment and Mitigation

The concept of fairness has been extensively researched in the context of centralized learning. Work on fairness can be categorized into two main areas: *bias assessment* (or *bias detection*) and *bias mitigation.* Bias assessment deals with finding biases in models using fairness metrics. Bias mitigation actively tries to diminish bias that arises within a model pipeline. This can be done using pre-processing (e.g. Kamiran and Calders, 2012; Calmon et al., 2017; Brunet et al., 2019; Wang et al., 2022), in-processing (e.g. Kamishima et al., 2012; Louizos et al., 2015; Berk et al., 2017; Wadsworth et al., 2018; Wu et al., 2018), or post-processing techniques (e.g. Hardt et al., 2016; Lohia et al., 2019; Lohia, 2021; Mehrabi et al., 2021a). Over the last few years, the interest of the combined field of privacy and fairness has increased (e.g. Ferraguig et al., 2021; Fioretto et al., 2022; Chang and Shokri, 2023; Shaham et al., 2023).

2.2 Privacy

Given recent developments in data privacy (e.g. GDPR law, EU AI ACT), an increased interest in privacy and security issues within in AI systems has lead to extensive research into privacy-preserving learning models. In the Cambridge Dictionary, the word privacy means *someone's right to keep their personal matters and relationships secret* (McCreary, 2009). Measuring and ensuring this concept can be complex. In AI, where machine learning models require vast amounts of data to function effectively, it is challenging to develop privacy-preserving methods that allow models to analyze and learn from data without extracting or revealing sensitive personal information. The goal is to leverage data effectively while maintaining strict privacy protections, ensuring that personal privacy is not violated in the process. There are different methods to achieve privacy (e.g. federated learning, cryptography and differential privacy), and all of these methods require a different view and approach for fairness research. This research focuses on federated learning, a promising technique for locally private training which is already applied in industry to tackle the issue of data silos and cross-platform privacy, while also maintaining good utility (Mammen, 2021). In the following subsections, all three commonly-used privacy-preserving methods will be briefly discussed, due to their overlapping nature and common goal of enhancing privacy.

2.2.1 Differential Privacy

Differential privacy (DP) provides a mathematical definition of privacy that ensures that a computational mechanism (e.g. an algorithm) will not reveal sensitive information of any individual in a dataset (Dwork et al., 2006). This means that, for a mechanism to be differentially private, the removal or addition of a single datasample will not affect the outcome of that mechanism, such that it is not possible to relate the outcome back to an individual (Dwork, 2008). There are different methods that can be used to ensure differential privacy in machine learning, e.g., adding random noise to the data during training (Dwork, 2008; He and Cai, 2017). Differential privacy is sometimes incorporated in federated learning pipelines for extra privacy protection (Banse et al., 2024).

To accommodate differentially private machine learning, Differential Privacy Stochastic Gradient Descent (DP-SGD) is commonly used as learning algorithm (Xie et al., 2021). Recent studies have shown that the negative performance impact of using DP-SGD over traditional SGD is exhibited more frequently for underrepresented groups (Bagdasaryan et al., 2019; Pujol et al., 2020; Tran et al., 2021). In these studies, the bias is measured as the disparity in performance between the predicted outcome on data with DP-noise and the ground truth. There is also theoretical work that shows that it is impossible to achieve differential privacy, fairness and reasonable accuracy at the same time (Cummings et al., 2019).

2.2.2 Cryptography

Another privacy-preserving method uses cryptography to make sure that the data remains private and secure throughout the machine learning process. A common cryptography method is homomorphic encryption, allowing for encrypted data on which certain operations can still be performed (Briguglio et al., 2021). Other methods rely on multi-party computation (MPC), which enables multiple parties to securely compute functions, ensuring both the correctness of the output and the privacy of all participants (Catalano et al., 2005). MPC algorithms use cryptography to guarantee these privacy and correctness protections. Unlike federated learning, MPC provides a higher level of security due to its use of intensive cryptographic operations. Both homomorphic encryption and multi-party computation are frequently used in combination with federated learning to enhance privacy protection (Byrd and Polychroniadou, 2020; Fang and Qian, 2021).

Existing fairness research on multi-party computations focuses on contribution fairness, i.e. whether all parties are always guaranteed an output (Choudhuri et al., 2017). This is different from group fairness, which considers fairness as an equal accuracy across demographic groups, and out of scope for this research. Homomorphic encryption is sometimes added to common fairness techniques to obtain privacy-preserving fair models (Shaham et al., 2023).

2.2.3 Federated Learning

The term *federated learning* was first introduced by Google in 2016 to describe their new approach for collaboratively learning to predict user's text input on tens of thousands Android devices without sharing personal information. Although the concept of distributing data and computations across multiple servers to accelerate AI training existed previously, applying this method to privacy-preserving decentralized learning was a novel idea at the time.

A federated learning framework consists of two main parts: a client side and a server side. The client side consists of various *clients* (or *parties*), each possessing their own data silo (i.e. a smaller isolated dataset), known as a *client dataset* or *local dataset*, on which they can locally train a model. On the server side, these trained local models are aggregated to form a global model that contains information from all clients. The distinction between client and server side is important, since client-specific data is available on the client side, but not on the server side, thereby maintaining local client privacy.

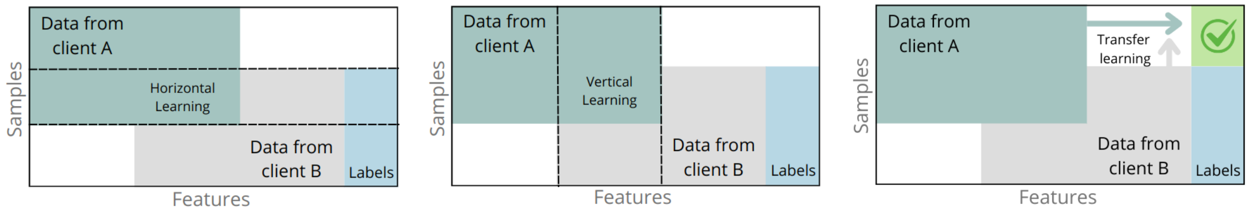
Federated learning is proposed and commonly-used as a privacy preserving solution for machine learning. However, recent work has shown that sensitive personal information can be inferred from the model parameters that are shared between clients during the federated learning pipeline, thereby compromising privacy (Carlini et al., 2019). The effectiveness of such privacy attacks is dependent with the complexity of the model and the client's dataset size. Proposed solutions for this problem include the addition of differential privacy (Gu et al., 2022) or data augmentation (Kaya and Dumitras, 2021). Despite these challenges, federated learning remains a vital approach for decentralized and privacy-aware machine learning, and it forms the basis for the methods used in this research.

A training pipeline for federated learning consists of four main steps. These steps are repeated over several iterative rounds.

1. **Client selection**: Picks the clients that will participate in that round.
2. **Parameter broadcasting**: The global model parameters are sent to the selected clients.
3. **Local model training**: The selected clients retrain the received global model on their local data.
4. **Model aggregation**: The clients send back their local model parameters. These will be aggregated to form the updated global model.

Types of Federated Learning

Federated learning can be categorized into several types based on the architecture and the specific applications for which it is employed. *Horizontal federated learning*, or vanilla federated learning, is the most common type, where the participating client datasets share the same feature space but have different individual data samples. This type is particularly useful in scenarios where the same type of data is collected across many users. A visualization of this type can be seen in Figure 2.1a. In *vertical federated learning*, different clients datasets share individual dataset samples, but hold different feature information for these samples. This type of federated learning is suitable for scenario's in which organizations might benefit from collaboratively leveraging their unique but complementary datasets. A visualization of this type can be seen in Figure 2.1b. Lastly, there is *federated transfer learning*, which extends the concept of federated learning by allowing for the transfer of knowledge between domains that have different feature spaces and data distributions. A visualization of this type can be seen in Figure 2.1c. Similarly to the current literature, this research will focus exclusively on horizontal federated learning.



(a) Horizontal Federated Learning (b) Vertical Federated Learning (c) Federated Transfer Learning

Figure 2.1: Visualization of data dimension overlap in different types of federated learning.

Aggregation methods

The goal of learning algorithms is to learn the model parameters that provide the best accuracy for a given task. This means that the optimization problem seeks to find the model parameters $\mathbf{w}$ that minimize the loss function $f(\mathbf{w})$. This is done by minimizing the average loss over all n training samples. This general optimization problem is more formally defined in Equation 2.1. The loss function $f(\mathbf{w})$ is obtained by summing over all loss functions $f_i(\mathbf{w})$ that represent the loss function for a specific data sample i.

$$\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w}) = \min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{w}) \tag{2.1}$$

In the context of federated learning, the n training samples are distributed across different clients, meaning that the computation will also be distributed across multiple clients. If the number of training samples for a client k is given as $n_k = |\mathcal{P}_k|$ (where $\mathcal{P}_k$ is the local training dataset of client k), then the objective function of $f(\mathbf{w})$ (Equation 2.1) can be written according to Equation 2.2. The parameters $\mathbf{w}$ are found by again minimizing the loss function $f(\mathbf{w})$, but this function is now calculated as a weighted average over all K client loss functions $F_k(\mathbf{w})$. Here, each client loss function $F_k(\mathbf{w})$ is calculated by summing over the loss of each data sample within the client dataset $\mathcal{P}_k$.

$$\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w}) = \min_{\mathbf{w}\in\mathbb{R}^d} \sum_{k=1}^{K} \frac{n_k}{n} F_k(\mathbf{w}) \qquad \text{where,} \quad F_k(\mathbf{w}) := \frac{1}{n_k} \sum_{i\in\mathcal{P}_k} f_i(\mathbf{w}) \tag{2.2}$$

Equation 2.2 shows that the weights of the global model are obtained by averaging over the weights of the local client models. There are different ways in which the local client models can be aggregated to update the global model. Different aggregation algorithms can differ in what they aggregate (e.g. model gradients or model weights), when they aggregate (synchronous or asynchronous) and whether local updates are combined with global model updates (Nilsson et al., 2018). Two commonly-used methods, FedSGD and FedAvg, will be discussed here.

FedSGD One simple aggregation technique is FedSGD (McMahan et al., 2017), inspired by traditional Stochastic Gradient Descent (SGD). In traditional centralized learning with SGD, the gradients are often computed on batches, which are subsets of the entire dataset. In federated learning, each client has a local dataset, analogous to these batches. In FedSGD, the central model is first distributed to all clients. Each client trains this model on their local dataset, thereby computing local gradients. Similarly to what is done with batches, the local gradients are then aggregated in proportion to the local dataset size of each client to calculate the gradient descent step in the central model.

Let's assume that w_t are the current global model parameters. With FedSGD the new global model parameters w_{t+1} can be obtained with a weighted average of all local client gradients g_k, as shown in Equation 2.3. The weights are defined by the size of the local dataset n_k. In this formula, we define the gradient for a client k as the gradient of the local client model $g_k = \nabla F_k(w_t)$ and $F_k(w_t)$ is the loss function of client k at time step t as defined in Equation 2.2.

$$w_{t+1} \leftarrow w_t - \eta \sum_{k=1}^{K} \frac{n_k}{n} g_k \tag{2.3}$$

FedAvg Instead of sharing the gradient updates with the central server, it is possible to share the local model weights. This is done for the FedAvg aggregation method (McMahan et al., 2017). FedAvg allows clients to do more than one gradient update locally, i.e. update the local model multiple times before sending it to the central server. This time, the local model weights (i.e. model parameters) will be aggregated in proportion to the local dataset size of each client to calculate the parameters of the global model.

The update step for FedAvg differs slightly to the one given in Equation 2.3 for FedSGD. Within FedAvg, the weights are first updated locally for each client k. This is done using the gradient from the local client model, as shown in Equation 2.4. Here, the gradient of client k is again defined as the gradient of the local client model $g_k = \nabla F_k(w_t)$ and η_k is the local learning rate of client k.

$$w_{t+1}^k \leftarrow w_t - \eta_k g_k \quad \forall k \tag{2.4}$$

The local client model parameters are then averaged together to obtain the updated global model parameters, as shown in Equation 2.5. This is done using a weighted average, in which the weights are again defined by the local client dataset size n_k.

$$w_{t+1} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} w_{t+1}^k \tag{2.5}$$

FedAvg is a generalization of FedSGD that, unlike FedSGD, has the advantage of allowing for multiple local client updates per iterative round. Furthermore, FedAvg has proven to achieve the highest accuracy among different federated learning algorithms (Nilsson et al., 2018). Therefore, this thesis will utilize FedAvg as the aggregation algorithm for federated learning.

Client Heterogeneity

A phenomenon increasingly studied in federated learning is that of client heterogeneity. In federated learning, often, the participating clients are deployed in different environments, resulting in divergence between local training conditions. There are different aspects of the local training that can lead to client heterogeneity.

The most evident cause of data heterogeneity is the data itself. *Data heterogeneity* can be caused by differences in label or feature distributions or by differences in quantity between different client datasets. An example of this phenomenon is the scenario in which the number of white people in one hospital is larger, while in another hospital the population of Hispanic people is much more present. In this case, the distribution of the feature race differs among different clients. Data heterogeneity can cause problems with respect to accuracy (Mammen, 2021) and fairness (Abay et al., 2021). Data heterogeneity can make optimization of a global model more difficult. Different clients train their local models on their local data, which leads to updates toward completely different local minima in the case of data heterogeneity. This local convergence point might not be aligned with the optimal objective for the global model. Such local updates can then drift the aggregated global optimal from the ideal global optimal, thereby overfitting to local clients. This results in non-optimal global convergence, thereby impacting the global accuracy.

Additionally, different clients may have varying computational resources. Client devices can vary in hardware and implementation, such as using different coding languages among different clients; for instance, some mobile devices use Java, while others are Python-based. This can result in local training processes that differ in implementation and execution. These differences fall under *computational heterogeneity*. It is important that these differences are taken into account within the federated learning pipeline, such that all clients are able to participate in the training process.

Federated learning requires communication between the local clients and the global server in the form of downloading and uploading model updates (e.g. model parameters or gradient updates). Since different clients can have different hardware and implementation, they can also have different approaches for communicating to the server-side. These differences can lead to differences in download and upload speeds. This is called *network heterogeneity*. This is also something that needs to be considered within the federated learning pipeline, such that all participating clients can update the global model correctly and efficiently.

Client Selection

Usually, only a fraction of the clients is selected every round to update the global model. This approach is used to enhance efficiency. In federated learning scenario's there is often a limited amount of communication and computation resources that make it impossible to let all clients participate every round (Li et al., 2022). Consequently, for every round, a careful selection of clients is crucial to obtain the best possible training results. There are roughly two approaches for the selection of clients: unbiased, where clients are selected at random or in proportion to their local dataset size, or biased, in which the clients are selected that lead to the fastest model convergence (Cho et al., 2020).

Despite the benefits of increased efficiency, client selection has its pitfalls. By only selecting a fraction of the clients, the effects of data heterogeneity can be aggravated, resulting in non-optimal convergence of the global model. Since the type of client selection can increase the impact of heterogeneity, it can also be a factor that gives rise to bias in a federated learning pipeline.

2.3 Fairness in Federated Learning

Previous research on fairness-aware federated learning can be broadly classified into seven distinct notions of fairness, as listed by Shi et al. (2023). Most of these fairness notions focus on achieving fairness across clients, thereby aiming for equal performance (*client fairness*, *good-intent fairness*), equal participation (*selection fairness*) or equal reward (*contribution fairness*, *regret distribution fairness*, *expectation fairness*). Additionally, they mention the notion of *group fairness*, which seeks to minimize disparities in algorithms decision-making across different demographic groups, and has found its way in the field of federated learning (Abay et al., 2020; Zhang et al., 2020; Du et al., 2021; Ezzeldin et al., 2023). This thesis focuses on group fairness, where fairness is measured using common fairness constraints (e.g. demographic parity, equalized odds) across different demographics. The following sections will provide a comprehensive overview of current work on group fairness in federated learning.

2.3.1 Causes of Bias in Federated Learning

Some previous work focuses on the causes of bias in federated learning (Abay et al., 2020, 2021). These causes can be roughly divided into four groups. First, bias can emerge from traditional causes that also lead to bias in centralized learning settings. These causes include, among others, prejudice, underestimation, and incorrect sampling or labeling (Kamishima et al., 2012). Secondly, bias can be caused by data heterogeneity. Data heterogeneity, i.e. structural differences in datasets between clients, affect how a global model is trained. This can lead to over- or underfitting with respect to certain demographics, and limits the effect of local debiasing efforts, thereby possibly resulting in a biased global model (Mendieta et al., 2022;

Ezzeldin et al., 2023). A third cause are the aggregation methods, which can cause bias, e.g. by the choice of aggregation weights or functions (Abay et al., 2020). Aggregation weights, for instance, can amplify the effect of over- or under-representation of sensitive groups in local datasets, which can lead to bias. Lastly, there is client selection and subsampling that can cause bias. Some FL algorithms do not query every client equally over all training rounds (Chai et al., 2020; Cho et al., 2020; Nishio and Yonetani, 2019). This unequal client selection can cause negative effects on sensitive groups when the selection choices are correlated to sensitive attributes.

2.3.2 Bias Assessment in Federated Learning

Within federated learning, the issue of measuring and reducing group fairness without accessing all data for sensitive attribute examination is still seen as an open research question (Kairouz et al., 2021). The privacy-preserving nature of federated learning makes it challenging to perform fairness research. In centralized settings, group fairness is measured by comparing the performances of demographic groups that are extracted from all data samples. This is hard to measure in federated learning because it is not possible to group individuals based on demographic attributes across the entire dataset. Each client's dataset remains local and private, which prevents the aggregation of all data necessary to form and analyze demographic groups across different clients.

The goal of bias assessment in federated learning is to measure bias in the global model while preserving privacy constraints. Research on bias assessment in federated learning is limited. In our literature research, we have found only one paper which actively evaluates bias within a federated learning framework (Chang and Shokri, 2023). Other studies on fairness-aware federated learning focus on mitigating bias in FL pipelines, using bias assessment techniques to evaluate their methods. Existing approaches can be categorized into three main groups: local bias assessment, aggregated local bias assessment and global bias assessment. The remainder of this section discusses these methods, and highlights their potentials and flaws.

Local Bias Assessment

With local bias assessment, the global model is evaluated on local client datasets, resulting in a bias assessment per client. This approach is adopted by Chang and Shokri (2023). Their work focuses on the local detection of bias, by measuring the impact of federated learning on the bias that arises on client-level.

The authors assess the impact of federated learning on fairness using a fairness gap between different training methods, measured through the metrics equalized odds (Hardt et al., 2016) and demographic parity (Dwork et al., 2012). They compare this fairness gap across three approaches: central learning, federated learning, and standalone local learning (i.e. a client trains a model locally using only their own client data). This comparison is done by evaluating all three models for accuracy and fairness on one local client dataset. In other words, the central[1] and global model will also be evaluated using the test set of the local client dataset that is used for training the standalone model. The authors show that centralized learning improves both accuracy and fairness compared to standalone learning, but that FL can exacerbate the bias found in the standalone models. Furthermore, the authors employ Integrated Gradient to measure the attribution of sensitive attributes to the models' predictions and thereby identify the sources of disparate treatment. The authors show that biased clients encode bias in a few model parameters (i.e. model weights) within their local models. This bias is propagated

[1] The model obtained from centralized training will be referred to as the 'central model' throughout the remainder of this thesis.

through the local model aggregation in federated learning, which results in a biased global FL model.

The bias assessment approach from Chang and Shokri (2023) primarily concentrates on bias that arises within individual clients' datasets. These local perspectives are crucial for understanding client-specific model behavior. Another advantage of this approach is that the local privacy of federated learning is preserved during the evaluation process. However, this approach only provides a bias measurement for each client, and thus does not provide insights in the perpetuation of bias within the global model across all clients.

Aggregated Local Bias Assessment

This approach measures bias on the client side, using the local client datasets to compute a bias score specific to each client. These local bias scores are then combined using a similar approach to the aggregation approach of the locally trained models, to obtain a global bias score. Equation 2.6 shows how the global bias measurement m_{global} is constructed from all local bias measurements m_k. Here, each client measurement m_k is obtained by evaluating the local model on their local dataset for a specific fairness metric. Furthermore, K is the total number of clients, n_k is the local dataset size and n corresponds to the global dataset size (i.e. a concatenation of all client datasets). In this case, the aggregation weight corresponds to the local client dataset size, however any aggregation weighting scheme can be used, as long as it is similar to the weighting scheme used during local model aggregation.

$$m_{global} = \sum_{k=1}^{K} \frac{n_k}{n} m_k \tag{2.6}$$

This method is used by Ezzeldin et al. (2023). In their work, they propose FairFed, a fairness-aware aggregation method that reweighs a clients' contribution based on the mismatch between the current global fairness measurement (m_{global}) and the local client fairness measurement (m_k). They did their experiments for different sensitive attribute distributions across clients to account for data heterogeneity. Their proposed method is evaluated by aggregating local bias scores with the same aggregation weights that are used for the model aggregation. They used demographic parity and equalized odds as fairness metrics.

A similar bias evaluation approach is used by Zhang et al. (2020). In their work, they propose FairFL, a federated learning framework that consists of a reinforcement learning framework for client selection and a secure information-aggregation protocol that optimizes fairness and accuracy. They evaluate their model by measuring a local discrimination index for each client. This discrimination index measures the bias towards a specific sensitive group (e.g. *female* or *Asian*). The local discrimination indices are aggregated using a similar aggregation method as is used for model aggregation. The authors implemented an additional security mechanism using polynomial functions to ensure that only the aggregated discrimination index can be reconstructed, thereby preventing the possibility to reconstruct local discrimination indices.

A big advantage of aggregating local bias is that it is possible to assess biases across all clients without losing privacy-preservation. Although this method shows a lot of potential, previous work has failed to provide any theoretical or experimental foundation to show the correctness of this approach. This thesis tries to address this shortcoming by providing an experimental analysis of this aggregated local bias assessment method.

Global Bias Assessment

A global bias assessment measures bias in federated learning models on a global test set that is a concatenation of all client test sets. This approach disregards the client partition and their local privacy.

Abay et al. (2020) uses two central bias mitigation methods, *reweighing* (Krasanakis et al., 2018; Blow et al., 2024) and *prejudice remover* (Kamishima et al., 2012), as inspiration for their federated learning bias mitigation methods. Reweighing adds weights to data samples of the training set. In federated learning, this can be done locally (on the local training sets), or globally by employing differential privacy. Prejudice remover can be employed in federated learning by adding a regularizer to every local model. The authors evaluate their method by measuring the bias with common fairness metrics (i.e. equalized odds, demographic parity) on the global test set, thereby not taking privacy into account during their testing phase.

Zeng et al. (2021) propose FedFB, a federated learning approach that is inspired by the centralized fair learning algorithm FairBatch (Roh et al., 2020), which incorporates this algorithm in FedAvg. They measure bias centrally on the global test set using demographic parity.

An advantage of this method is that it is an easy and correct implementation to assess bias globally across all clients. However, a disadvantage of this method is that it fails to preserve the privacy constraint of federated learning. To account for this issue, there is work that combines a global bias assessment with differential privacy (Liu et al., 2022), although this brings additional complications (Garfinkel et al., 2018).

2.3.3 Bias from Data Heterogeneity

Even though Abay et al. (2020) mention data heterogeneity as a possible cause for unfairness in federated learning, the work that considers data heterogeneity in the context of fairness in federated learning is still limited. Abay et al. (2021) were first in assessing the performance of their bias mitigation techniques (reweighing and prejudice remover) under similar data distributions (IID) and under highly imbalanced data distributions (non-IID), thereby accounting for data heterogeneity across clients. They simulated this scenario using two clients that contained a ratio of 85-15, 99-1 or 100-0, on binary sensitive attributes (sex; male and female, and race; white and black). They found that FL models that are trained over highly imbalanced data distributions may learn bias from clients in an imbalanced manner.

Ezzeldin et al. (2023) evaluate their approach on diverse sensitive attribute distributions by exploiting a generic non-IID synthesis method based on Dirichlet distribution controlled by a hyperparameter α, such that $\alpha \to \infty$ results in complete IID distributions. They found that their proposed FairFed method outperformed previous local debiasing methods under heterogeneous distributions, but failed to do so under homogeneous data distributions.

2.4 Research Gaps

Given this previous work on fairness-aware federated learning, there are two major gaps that can be identified in current literature. Some of the proposed bias assessment methods are promising, particularly the aggregated local bias assessment, because it preserves the private nature of federated learning, while also allowing for a comprehensive bias assessment across all clients. Other methods do not simultaneously offer these advantages without relying on differential privacy. This is not always desirable, since incorporating differential privacy can introduce additional complexity and potentially reduce utility (Garfinkel et al., 2018).

Despite its advantages, the aggregated local bias assessment method lacks a thorough evaluation of its accuracy. This method assumes that bias can be aggregated in the same way as model parameters, implying that local bias measurements can be combined to reflect global bias accurately. However, model parameters differ significantly in dimensions, structure, and purpose, raising doubts about the reliability of this approach for measuring global bias, especially since it has not been empirically, theoretically nor analytically verified. In other words, it is uncertain whether aggregating local bias scores yields the same bias score as a global bias assessment, making it a potentially unreliable privacy-preserving alternative.

The lack of an accurate, privacy-preserving bias assessment methodology for federated learning also led to an absence of research that provides a global evaluation of bias that can emerge during the federated learning pipeline. Chang and Shokri (2023) demonstrate bias propagation on client level, but previous studies have not assessed this on a global scale. Furthermore, the degree of data heterogeneity among clients is expected to influence the bias that arises in federated learning, as noted by Abay et al. (2020) and Abay et al. (2021).

This thesis tries to address these shortcomings by providing a systematic analysis on bias in federated learning and how this can be assessed. The contributions of this research are as follows:

- This work evaluates the aggregated local bias assessment method that is currently used in federated learning by comparing it to a global bias assessment.
- This work measures global bias that arises from the federated learning framework and compares this to bias that arises from a central learning and evaluation pipeline.
- Both these assessments are conducted across different types and amounts of data heterogeneities.

The limited research on the aggregated local bias assessment methodology makes it difficult to hypothesize about the outcome of the first research objective. However, only one example of a discrepancy between the global assessment and the aggregated local assessment is enough to dispute the reliability of the aggregated local assessment. The lack of theoretical backup of this approach lead us to expect that such a discrepancy will be identified. Furthermore, it is expected that this divergence can be influenced by the type and amount of data heterogeneity. Similarly to the local assessment results from Chang and Shokri (2023), it is expected that an FL model shows more bias compared to a centralized learning model. Since previous research has indicated that data heterogeneity between clients can affect the performance and fairness of federated learning models it is expected that the introduction of data heterogeneity will exacerbate the bias in the federated learning model when compared to a centralized model.

Chapter 3

Methodology

This research focuses on bias assessment in federated learning with two main objectives: (1) evaluating the accuracy of the bias assessment approach that aggregates local bias measurements for a global fairness evaluation, and (2) assessing whether the federated learning pipeline itself introduces biases. Experiments were designed and conducted to address both research objectives using three training and testing pipelines.

3.1 Training and Evaluation Pipelines

Three training and testing pipelines were established for fairness evaluation, enabling the comparison needed to address both research questions. An overview of the experimental setup is given in Figure 3.1. To assess the correctness of the *aggregated local bias assessment* method, a federated learning pipeline that uses local bias detection on a global FL model (i.e. the model that consists of all aggregated local models) and aggregates the locally obtained bias scores together (Pipeline 1), will be compared to a federated learning pipeline that uses global bias assessment on the global FL model (Pipeline 2). This comparison will answer the first research question. To measure the bias that arises in the federated learning pipeline, a federated learning pipeline that uses a global bias assessment on the final global FL model (Pipeline 2) is compared to a centralized learning pipeline (Pipeline 3), thereby answering the second research question. The training objective for all three pipelines is a simple classification task, for which a Logistic Regression model will be employed.

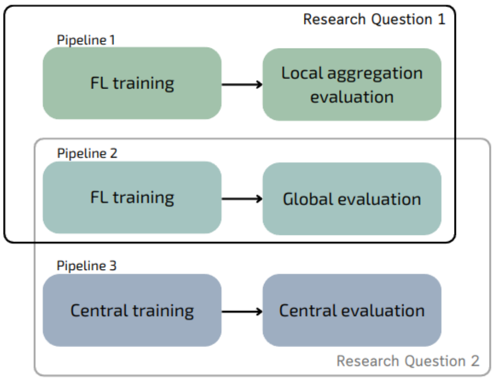


Figure 3.1: Overview of experimental setup of all three pipelines and how they will answer both research questions.

Pipeline 1: Federated Learning with Aggregated Local Bias Assessment

The first pipeline[1] uses federated learning for training and employs an aggregated local bias assessment. Within this pipeline, the data is first split up in different clients. These different clients train their local models on their local datasets. These locally trained models are aggregated using a FedAvg algorithm (see Section 2.2.3) to obtain the global FL model. This is done in a few rounds of local retraining and aggregation. After the last training round, the bias is assessed for every client on the last updated global model using the local client dataset to obtain a *local bias score* for every client. This score is called a local bias score, but represents the bias that arises in the *global* model with respect to a *local* dataset. The local bias scores of all clients are then aggregated together to obtain a *global bias score*. This is done using the same aggregation technique as the FedAvg method for model parameters; taking a weighted average of all local bias scores, weighted by the dataset size per client. The formula by which this is done is identical to Equation 2.6. A visualization of this pipeline can be found in Figure 3.2.

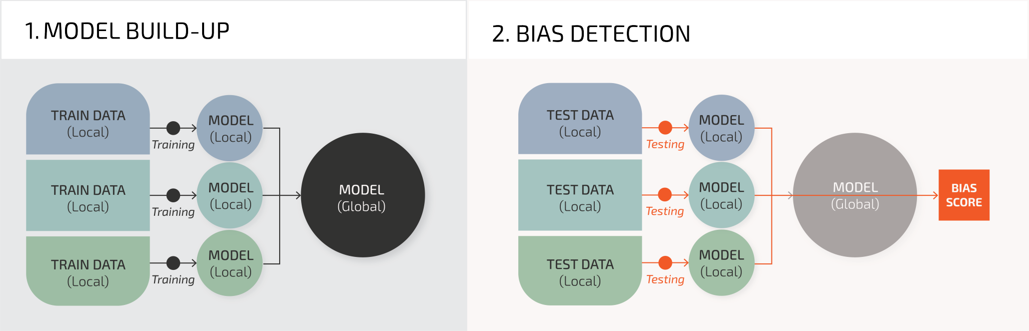


Figure 3.2: Overview of local pipeline: federated learning pipeline with aggregated local bias assessment.

Pipeline 2: Federated Learning with Global Bias Assessment

The second pipeline[2] uses federated learning for training and employs a global bias assessment on the trained global FL model. Within this pipeline, the data is first split up in different clients. These different clients train their local models on their local datasets, after which the local models are aggregated using a FedAvg algorithm. This local training and aggregation is again repeated for a few rounds to obtain the global FL model. This pipeline assesses bias by first concatenating all local client datasets into a global dataset. The bias in the trained global model is then measured using this global dataset. A visualization of this pipeline can be found in Figure 3.3.

Pipeline 3: Central Learning

The third pipeline[3] employs a traditional centralized learning approach for training and bias assessment. The model is trained on a global dataset without splitting the data among different clients. Bias in the centrally trained model is then measured using the test samples of the whole dataset. A visualization of this pipeline can be found in Figure 3.4.

[1] In the remainder of this thesis, this pipeline will be referred to as the local pipeline.

[2] In the remainder of this thesis, this pipeline will be referred to as the global pipeline.

[3] In the remainder of this thesis, this pipeline will be referred to as the central pipeline.

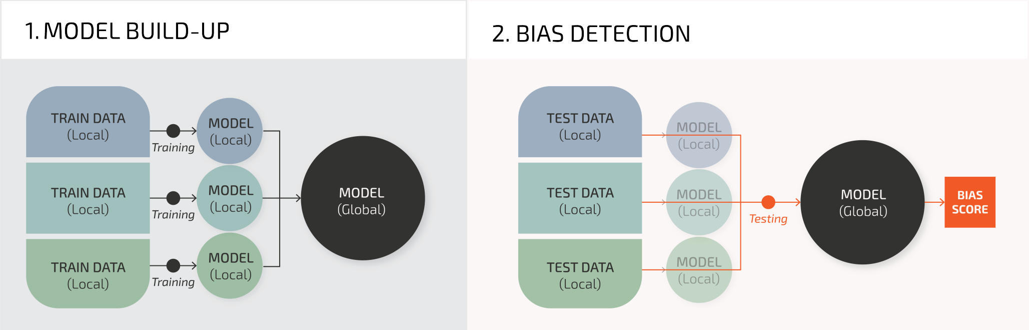


Figure 3.3: Overview of global pipeline: federated learning pipeline with global bias assessment.

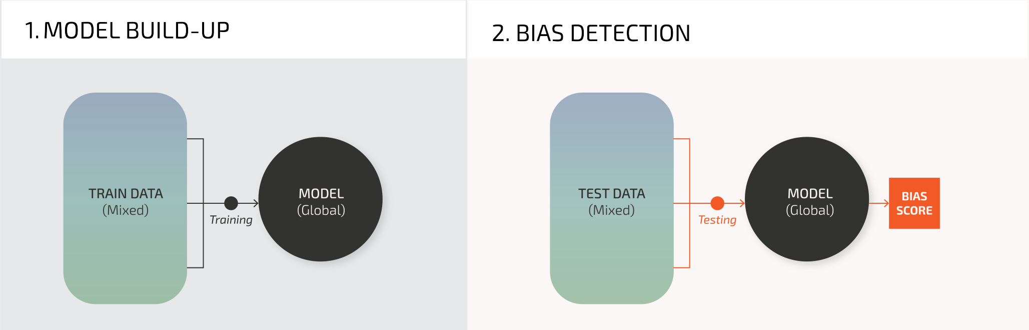


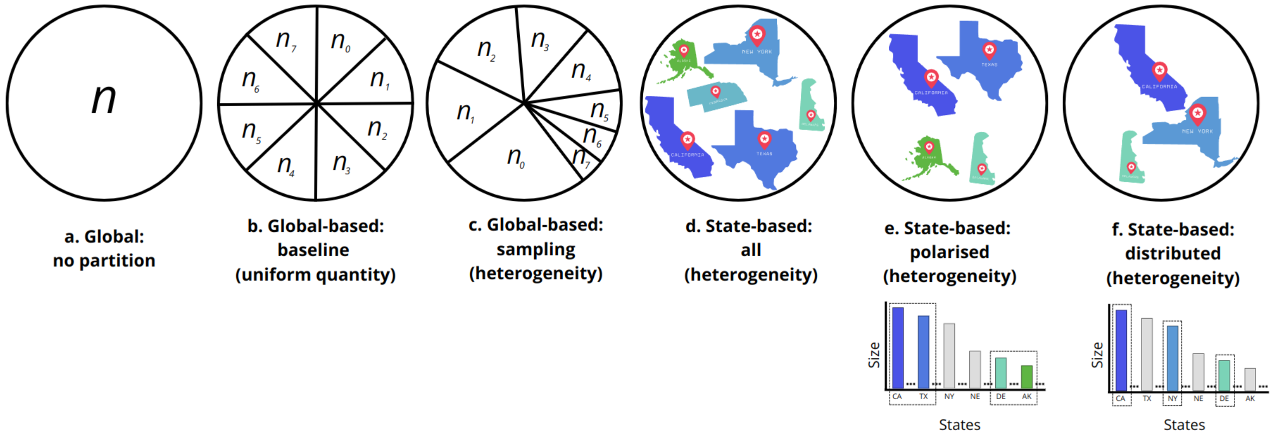
Figure 3.4: Overview of central pipeline: centralized learning pipeline with a centralized bias assessment.

3.2 Data Partitioning

The three pipelines will be tested across different experimental client partitions. To introduce the methodology for creating these client partitions, it is important to first understand the structure of the dataset used in this research. The dataset should possess some specific characteristics: it should consist of data samples that contain personal information, including sensitive attributes (e.g., race, age, gender) and socio-economic variables (e.g., education level) that relate to a general societal concept (e.g., income) that can be learned by a model. Additionally, the current approach to client partitioning requires data on the geographic location of each individual. Specifically, some proposed experimental client partitions are based on US states, with each state treated as a separate client. Therefore, the dataset structure must accommodate this geographical component.

3.2.1 Data Heterogeneity

To investigate the influence of the type and amount of data heterogeneity on the different pipelines, experiments will be conducted with various types and levels of data heterogeneity introduced into the federated client partition. There are three different components in which data heterogeneity can occur: (1) quantity, (2) label and (3) feature. Quantitative heterogeneity refers to differences in the sizes of local datasets among clients. Label heterogeneity indicates variations in label distributions between clients. Feature heterogeneity involves differences in feature distributions, potentially affecting one or multiple features across clients. These types of data heterogeneity, along with their combinations, will be artificially introduced by partitioning the dataset through deliberate sampling of US states and/or data samples from the complete global dataset. These additional experiments will be reported and compared to assess how data heterogeneity influences bias assessment within federated learning.



a. Global: no partition

b. Global-based: baseline (uniform quantity)

c. Global-based: sampling (heterogeneity)

d. State-based: all (heterogeneity)

e. State-based: polarised (heterogeneity)

f. State-based: distributed (heterogeneity)

Figure 3.5: Client partitions with homogeneity (b) or heterogeneity (c, d, e, f) in quantity.

Baseline

A baseline dataset partition should have equal local dataset quantities and identical label and feature distributions. This partition is created by randomly splitting the global dataset (Figure 3.5a) into equally sized client datasets (Figure 3.5b). Random sampling aims to create client datasets with label and feature distributions that match those of the global dataset. Although exact matching is not guaranteed due to sampling randomness, sufficiently large client datasets can approximate this balance.

Quantity Heterogeneity

Within this category, the client partition should have no heterogeneity in labels and features, but only have a difference in local dataset quantities. Considering the geographical nature of the dataset, which includes participants from all US states, there are two ways to create partitions with different local dataset quantities: global-based and state-based. The global-based approach samples individual data samples from the global dataset to obtain client datasets of varying sizes (see Figure 3.5c). This approach disregards the natural geographic partition that occurs within the original dataset. By randomly sampling individuals, it is assumed that the clients will have similar label and feature distributions.

The state-based approach samples states of different sizes. There are three ways in which a state-based partition with quantity heterogeneity can be created. First, the original state partition can be used, as states naturally have different population sizes reflected in the dataset. This method is shown in Figure 3.5d. Another method is to create a partition by ordering all states based on their dataset size and selecting some of the largest and smallest states to form polarized client sizes (Figure 3.5e). This method thus only uses a polarized subsection of all original US states. Given the order of states from big to small, it is also possible to sample the states on an interval to obtain states that are as much distributed as possible with respect to their size. This method thus uses a distributed subsection of all original US states, as shown in Figure 3.5f. In the case of sampling US states as clients, it is possible that different clients have a natural deviation in feature distributions. This can be seen as a limitation of this state sampling approach, thereby not guaranteeing complete homogeneity in feature and/or label distributions across clients. However, to obtain knowledge about a scenario in which quantity heterogeneity occurs without harming the natural geographic characteristics of a dataset, these approaches will still be used. This limitation will be taken into account when analyzing the result.

Label Heterogeneity

Within this category, the client partition should have label heterogeneity, but equal quantity between clients and preferably no feature heterogeneity. To establish such a partition, individual data samples will be categorized from the global dataset based on their label class, i.e. all data samples with a positive label are categorized together, and all data samples with a negative label are categorized together. These distinctive label categories can be divided further to obtain equally-sized client datasets. A downside to this approach is that the categorization of individual data samples for label heterogeneity can also automatically lead to feature heterogeneity between the obtained categories. If there is a strong relationship between one of the labels and a specific feature, the categorization of the data samples based on that label will automatically lead to a heterogeneous categorization in the related feature classes. This means that this approach cannot guarantee an absence of feature heterogeneity between different clients. An alternative approach would be to relabel data samples without changing their feature values. However, this would break the natural relationship between features and labels, which would negatively influence the training performance. Therefore, it is chosen to create the client partition with label heterogeneity by form of categorization of the original data samples. This can either be done for complete heterogeneity, where one client has only one label type (100-0), or for a heterogeneity where a client has a specific distribution of labels, such that 75% of each client dataset has one label, while the other 25% of the client dataset has the other label (75-25).

Feature Heterogeneity

Within this category, the client partition should have equal dataset quantities between clients, and preferably no label heterogeneity. The approach is similar to the label heterogeneity, where individual data samples are categorized based on their class for a specific feature. These categories are then used to create client datasets with feature heterogeneity. Similarly to label heterogeneity, it is now possible that the obtained client datasets automatically show varying label distributions, thereby accidentally introducing label heterogeneity alongside the desired feature heterogeneity. But again, breaking the relationship between features and labels is undesirable, such that for these experiments the potential additional label heterogeneity will be accepted. An additional downside to the categorization approach for feature heterogeneity is that it requires equal-sized client datasets. If one feature class has significantly fewer data samples than the other classes, larger feature classes must be reduced to match the smallest class size to obtain equally-sized clients (see partition (A) in Figure 3.6). This process will for the remainder of this paper be referred to as 'cut-off'. If the size difference between feature classes is substantial, this means that a lot of data will be disregarded, which eventually can influence the performance of the model. For features that already show equally sized classes, this cut-off is not necessary (see partition (A) in Figure 3.7). Non-binary features will only be tested with complete heterogeneity, since it is impossible to create clients with a consistent partial heterogeneity across all classes for multiple non-evenly distributed classes. Binary features will be tested on a partition with full heterogeneity across clients (100-0) and a partition with partial heterogeneity across clients (75-25).

Multiple Heterogeneities

It is also possible to create client partitions with a combination of different heterogeneities. Additional experimental partitions are set up to explore a combination of label heterogeneity and quantity heterogeneity, and a combination of feature heterogeneity and quantity heterogeneity. These experimental partitions use the same categorization method as is used for heterogeneity

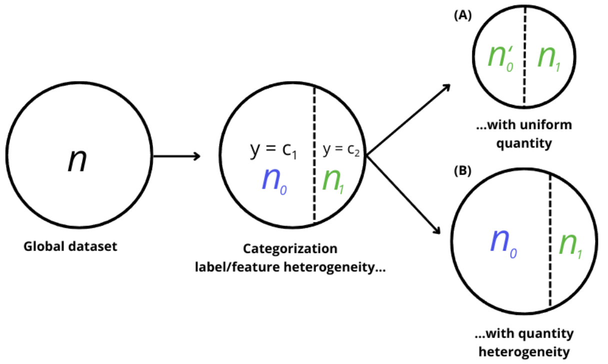


Figure 3.6: Obtain clients with uniform (A) or heterogeneous (B) quantity from heterogeneous categorization based on label/feature heterogeneity by disregarding labels in the case of uniform quantity.

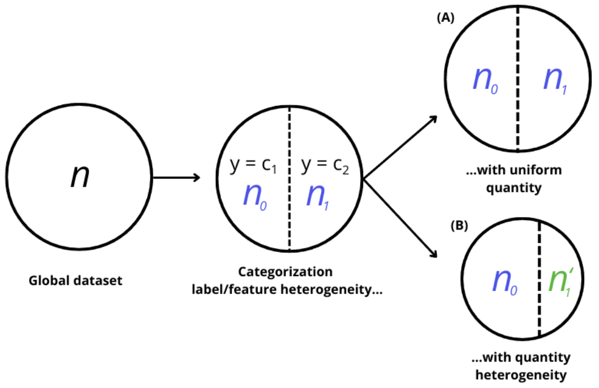


Figure 3.7: Obtain clients with uniform (A) or heterogeneous (B) quantity from homogeneous categorization based on label/feature heterogeneity by disregarding labels in the case of quantity heterogeneity.

in label and feature. Now, instead of cutting-off data samples to obtain equal client sizes, data samples are cut-off to obtain non-equally sized clients. This approach will thus disregard more data samples if the original categorization created relatively equal categories (see partition (B) in Figure 3.7). If the initial feature/label categorization lead to non-equally sized categories, this cut-off is not necessary (see partition (B) in Figure 3.6). Additionally, it is possible to create partitions based on a combination of heterogeneity in label and feature, or even on a combination of feature, label and quantity heterogeneity, by creating partitions based on combinations of feature classes and labels, and cutting-off data samples to obtain homogeneous or heterogeneous quantity partitions. However, categorizations of a combination of feature classes and labels can become very small, thereby leading to very small client datasets, which could negatively affect the training performance. It is therefore chosen to disregard these experiments in this study.

3.3 Bias Assessment

Following Ezzeldin et al. (2023), Abay et al. (2020), and Chang and Shokri (2023), bias detection is performed using two common fairness metrics: demographic parity (DP) and equalized odds (EO). These metric ratios provide a comprehensive overview of the *relative* performance disparities between demographic groups, facilitating clear comparisons between the different pipelines. The DP ratio requires that a prediction is done independently of the sensitive attribute group. This ratio is calculated as the ratio between the expected values of the lowest and highest performing groups. This is shown in Equation 3.1, where a is a sensitive attribute class (e.g. *White*) for a sensitive feature A (e.g. race). The input data samples X are processed by the classifier $h(\cdot)$ to compute the expected value across all sensitive attribute classes.

$$\text{DP ratio} = \frac{\min_a \mathbb{E}[h(X)|A = a]}{\max_a \mathbb{E}[h(X)|A = a]} \tag{3.1}$$

The EO ratio requires equal true positive (TPR) and false positive rates (FPR) across demographic groups, and is calculated as the minimum value between the TPR and FPR of the best and worst performing groups. This is shown in Equation 3.2, where the same variables are used as in the formula for DP ratio, with the addition of Y, which are the true labels for the input data samples X, to calculate the TPR and FPR across all sensitive attribute classes. For both metrics, a bias score of 1 indicates a fair model, while a score of 0 indicates an unfair model.

$$\text{EO ratio} = \min\left(\frac{\min_a P[h(X)|A=a,Y=1]}{\max_a P[h(X)|A=a,Y=1]}, \frac{\min_a P[h(X)|A=a,Y=0]}{\max_a P[h(X)|A=a,Y=0]}\right) \tag{3.2}$$

3.3.1 Sensitive attributes

This research will conduct its bias assessment on the sensitive attributes Sex and Race. The Sex[4] attribute, being binary, will compare *male* and *female* groups. The Race attribute, with multiple categories, will have its score based on the ratio between the best and worst performing race groups. Additionally, to create more robust race groups in which all protected and unprotected groups are considered, this attribute will additionally be converted to binary categories. Specifically, the protected group *white* will be compared to the unprotected group *non-white* (creating a White/Non-white categorization), and the unprotected group *black* will be compared to the protected *non-black* group (creating a Black/Non-black categorization), as *white* and *black* are the largest original race groups. The scores for these sensitive attribute categorizations are calculated such that the ratio is always *non-white* divided by *white*, and the group *black* divided by *non-black*, ensuring that the unprotected group is divided by the protected group. This allows for a consistent comparison between different experimental partitions and between the two metrics themselves.

Table 3.1: Overview of all bias metrics for all sensitive attribute categorizations considered during this research.

Metric	**Explanation**
EO Sex	Equalized odds score with respect to binary Sex-categories.
EO Race	Equalized odds score with respect to all Race-categories.
EO White	Equalized odds score with respect to binary Race-categories: *white*, *non-white*.
EO Black	Equalized odds score with respect to binary Race-categories: *black*, *non-black*.
DP Sex	Demographic parity score with respect to binary Sex-categories.
DP Race	Demographic parity score race with respect to all Race-categories.
DP White	Demographic parity score with respect to binary Race-categories: *white*, *non-white*.
DP Black	Demographic parity with respect to binary Race-categories: *black*, *non-black*.

[4] The attribute Sex is used to stay consistent with the phrasing of the used dataset. This definition does not consider genders, only the biological sex of an individual.

Chapter 4

Experiments

This chapter discusses the experimental setup that is used to evaluate the three training and evaluation pipelines. The code for all experiments is publicly available[1]. All pipelines are trained and evaluated on a single `NVIDIA A100-SXM4-40GB` GPU over five runs with different random seeds.

4.1 Dataset

For our experiments, we looked for a dataset meeting the requirements that it (1) contains data samples of individuals and personal information on their sensitive attributes, (2) has a learnable pattern between labels and features, and (3) is easily accessible. To evaluate the approaches in a real-world setting, it was also valuable to use a dataset with a natural partition suitable for client partitioning. The initial focus was on identifying a medical dataset, given the significant relevance of federated learning in the medical field. However, despite thorough research, no medical dataset met all the necessary criteria. The medical datasets that were considered but ultimately rejected are listed in Table A.1 in the appendix. Consequently, the search was expanded to include nonmedical datasets, leading to the selection of the American Community Survey (ACS) Public Use Microdata Sample (PUMS) dataset. The ACS PUMS dataset (Ding et al., 2021) forms an alternative to the more commonly-used *UCI Adult dataset*, and overcomes issues such as the age, limited documentation, and outdated feature encodings of the Adult dataset. The ACS PUMS datasets are created from American Census surveys from 2005 to 2022 and contain either information on individuals (person record) or on households (housing record). PUMS files either cover data over one individual year, or over a five-year period. More information on the dataset, and how it is extracted, can be found in Appendix A.

4.1.1 Prediction Task

There are different classification objectives that can be used from the UCS PUMS dataset. This research will consider the income prediction task, whereby the model is trained to predict whether an individuals' income is above or below 50,000 US dollars. This is done with a filtered version of the entire ACS PUMS dataset, called the ACS Income dataset, which only contains data samples of individuals above the age of 16, who reported usual working hours of at least 1 hour per week in the past year, and have an income of at least 100 US dollars. These preprocessing steps are equivalent to the filters used for the UCI Adult dataset (Kohavi and Becker, 1996).

[1] `https://github.com/jelkejm/experimental_analysis_of_bias_assessment_in_FL`

The ACS Income dataset contains personal information (e.g. sex, race, marital status) as well as professional information (e.g. occupation, educational level) of 1,720,154 American citizens. The classification task considers a total of nine features (an extensive list can be found in Appendix A). This dataset includes information on the US state where each individual resides, which can be exploited for a natural client partition. Each state can be treated as a separate client, with its own local dataset comprising all individual data samples from that state. The dataset is downloaded from `folktables`[2], a Python library that provides access to US Census datasets.

4.1.2 Data Analysis

The dataset is evenly distributed with positive data samples (individuals with an income above $50k) and negative data samples (individuals with an income below $50k). The percentage of male and female individuals is also evenly distributed across all US states. The distribution between *whites* and *non-white* is also relatively balanced. However, the classes of Race and Black/Non-black show disproportionate distributions, as seen in Figure 4.1. A further data analysis has shown that within the *male* and *female* groups, there is a slightly higher percentage of men with an income above $50k compared to women. This analysis also showed that the percentage of non-white individuals with an income above $50k is much smaller compared to the percentage of white individuals with an income above $50k. The details of this data analysis can be found in Appendix A. The specific ACS Income dataset that is used for this research contains personal data over only the year 2022 (i.e. `person`-record, 1-year, 2022) as it contains the most recent trends.

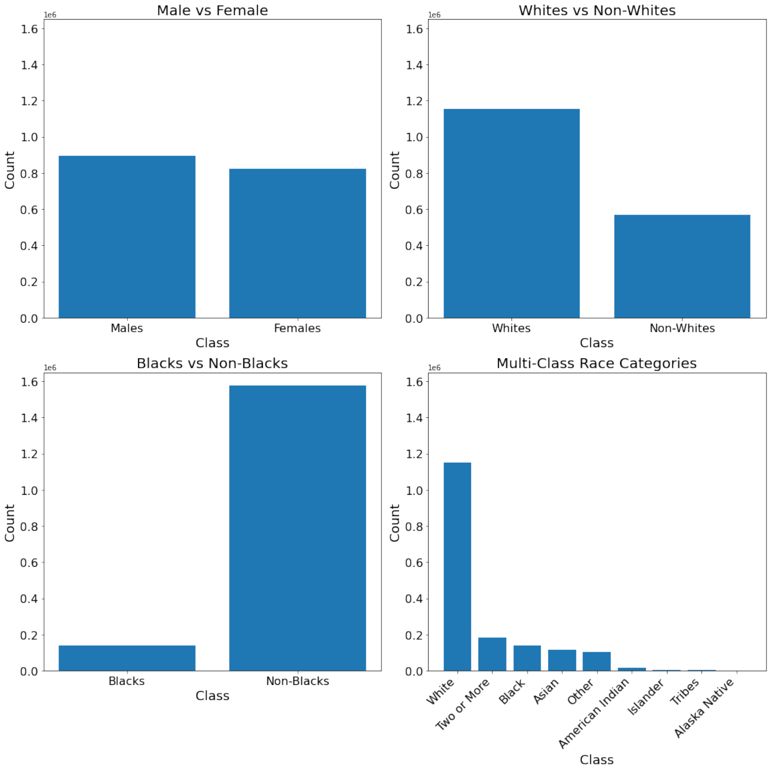


Figure 4.1: Distribution of different sensitive attribute classes for each sensitive attribute categorization.

[2] `https://github.com/socialfoundations/folktables/tree/main`

4.2 Federated Learning

The federated learning pipeline is implemented using the Flower framework. The aggregation algorithm that is used for these experiments is FedAvg. The global model that will be trained for income prediction is a Logistic Regression model. The local training is done using an Adam optimizer, a local learning rate of 0.001 and a batch size of 32. Every client trains the global model locally for one epoch, before sending the updated local parameters back to the server. These rounds of local training and model aggregation are repeated five times. Instead of employing a client selection algorithm, all clients are sampled for both the training phase and the evaluation phase. This allows for a good comparison with the centralized learning method that also uses all data for training and evaluation. Each client dataset undergoes an 80-10-10 split into training, validation, and test sets, where the train set was used for training, the validation set was used for hyperparameter tuning and the test set to obtain the final results.

4.2.1 Experimental partitions

This thesis will consider different client partitions in which heterogeneity in quantity, label and feature is introduced as proposed in Section 3.2. For this research, only a few important features, both sensitive and non-sensitive, were chosen to probe the influence of feature heterogeneity on bias detection. The chosen features are Race, Sex and Marital Status. The Race and Sex feature can give insights whether heterogeneity in sensitive attributes will also result in more biased results. The feature Marital Status is chosen as a non-sensitive feature because the number of disregarded samples (from the cut-off) is smallest for this feature (see Table B.1 in the appendix). The Race and Marital Status features will be only tested with complete heterogeneity, since these features entail multiple non-evenly distributed classes. The Sex feature will also be tested on a partition with full heterogeneity across clients (100-0) and a partition with partial heterogeneity across clients (75-25). This setup results into 17 different experiments, which include a baseline and 16 additional experimental client partitions. All client partitions are used for training and evaluation of all three pipelines. An overview of the different experimental partitions is given in Table 4.1. Every experimental partition is implemented to partition the data into 51 clients. Due to differences in heterogeneity types, the global datasets of different experimental partitions can differ in size, as can be seen in Figure B.

4.3 Central

The model that is employed for the income prediction task is Logistic Regression. For the centralized learning setting, the train-test-validation split of every client is concatenated to form the dataset used for central training and evaluation. This approach ensures that the train-test-validation split for centralized learning encompasses the same data samples as the combined distributed local client train-test-validation splits. The model is trained using an Adam optimizer, a learning rate of 0.001 and a batch size of 32. The model is trained for five epochs, to align with the five rounds (and single local epoch) of federated learning.

Table 4.1: All experimental partitions from this research and corresponding types of heterogeneities and how they are introduced. A star (⋆) indicates that the experimental partition has no heterogeneity with respect to that component.

Experimental partition	**Quantity**	**Label**	**Feature**
1 (Baseline)	⋆	⋆	⋆
2	Sampling	⋆	⋆
3	All	⋆	⋆
4	Polarized	⋆	⋆
5	Distributed	⋆	⋆
6	⋆	Income (100-0)	⋆
7	⋆	Income (75-25)	⋆
8	⋆	⋆	Sex (100-0)
9	⋆	⋆	Sex (75-25)
10	⋆	⋆	Race
11	⋆	⋆	Marital Status
12	Cut-off	Income (100-0)	⋆
13	Cut-off	Income (75-25)	⋆
14	Cut-off	⋆	Sex (100-0)
15	Cut-off	⋆	Sex (75-25)
16	Cut-off	⋆	Race
17	Cut-off	⋆	Marital Status

Chapter 5

Results

This chapter presents the results from the evaluation of the three model pipelines across various data partitions. The first research question investigates the accuracy of the aggregated local bias assessment by comparing the local and global pipelines. The second research question examines the influence of the federated learning pipeline on bias by comparing the global and central pipelines.

5.1 Evaluation of Aggregated Local Bias Assessment

The first experiments tried to answer the research question: *How accurate is the aggregation bias assessment method in detecting bias that arises in a federated learning framework?* For this, the bias scores from a trained FL model as measured with an aggregated local bias assessment and a global bias assessment are reported. The global bias scores are subtracted from the local bias scores to obtain an absolute metric difference, indicating whether the aggregated local assessment measures more bias (negative difference) or less bias (positive difference). This comparison is made across both the demographic parity (DP) and equalized odds (EO) metrics for the different sensitive attribute categorizations.

5.1.1 Baseline

The metric scores for the baseline data partition are presented in Table 5.1 as the average results over five runs with different random seeds, and their corresponding standard deviations. These results show similar bias scores for most metrics between the global and aggregated local assessment. However, the locally obtained scores for demographic parity and equalized odds for the attribute Race and the equalized odds score for Black/Non-black are significantly lower compared to their global counterparts, meaning that more bias is detected by the aggregated local method for these metrics.

5.1.2 Heterogeneity

The comparison between the global and aggregated local bias assessment is also conducted for different heterogeneous client partitions. Figure 5.1 show the absolute differences between the global and aggregated local bias assessment for all experimental partitions. The average differences across five runs with different random seeds are reported with their corresponding standard deviations.

Some values are missing across different partitions. For the partition with complete heterogeneity in label, the equalized odds metrics are missing since every client contains only one label class, causing that either the true positive rate or the false positive rate equals 0 across

Table 5.1: Bias metric comparison between global bias assessment (global) and aggregated local bias assessment (local) for baseline data partition as the average over five random seeds, and the corresponding standard deviations.

	Global	**Local**	**Difference**
Accuracy	0.735 (±0.001)	0.735 (±0.001)	0.000 (±0.001)
DP			
Sex	0.668 (±0.034)	0.669 (±0.034)	+0.001 (±0.048)
Race	0.347 (±0.010)	0.063 (±0.015)	-0.284 (±0.018)
White	0.592 (±0.006)	0.593 (±0.006)	+0.001 (±0.008)
Black	0.679 (±0.014)	0.679 (±0.013)	0.000 (±0.019)
EO			
Sex	0.711 (±0.060)	0.709 (±0.055)	-0.002 (±0.081)
Race	0.290 (±0.096)	0.001 (±0.002)	-0.289 (±0.096)
White	0.523 (±0.013)	0.525 (±0.013)	+0.002 (±0.018)
Black	0.779 (±0.013)	0.724 (±0.020)	-0.055 (±0.024)

all groups for each client. This complication also occurs in the data partition based on the combination of label and quantity heterogeneity. Similarly, in the case of feature heterogeneity in Sex or Race, the local metrics are disregarded since every client only has one sex or race class such that minimal and maximal values come from the same class, making the local ratio's nonsensical.

In Figure 5.1, both the DP and EO scores for the Race attribute show consistently larger differences across all experimental data partitions, similar to the baseline results. This pattern is also observed for the EO score of Black/Non-black, albeit with a smaller difference.

5.2 Bias in Federated Learning vs. Central Learning

The second objective of this research tried to address the question: *What bias arises during the federated learning process compared to a centralized learning process?* For this, the bias from a global FL model as measured with a global bias evaluation is compared to the bias as measured in a centrally trained model. The global bias scores are subtracted from the local bias scores to obtain an absolute metric difference, indicating whether the federated learning model induces more bias (negative difference) or less bias (positive difference) compared to centralized training, across both EO and DP metrics for different sensitive attributes.

5.2.1 Baseline

The metric scores for the baseline data partition are presented in Table 5.2 as the average results over five runs with different random seeds, and their corresponding standard deviations. These results show that federated learning induces more bias for the sensitive attributes Race, White/Non-white and Black/Non-black. But that with respect to Sex, the federated learning model is more fair. These tendencies are measured by both demographic parity and equalized odds.

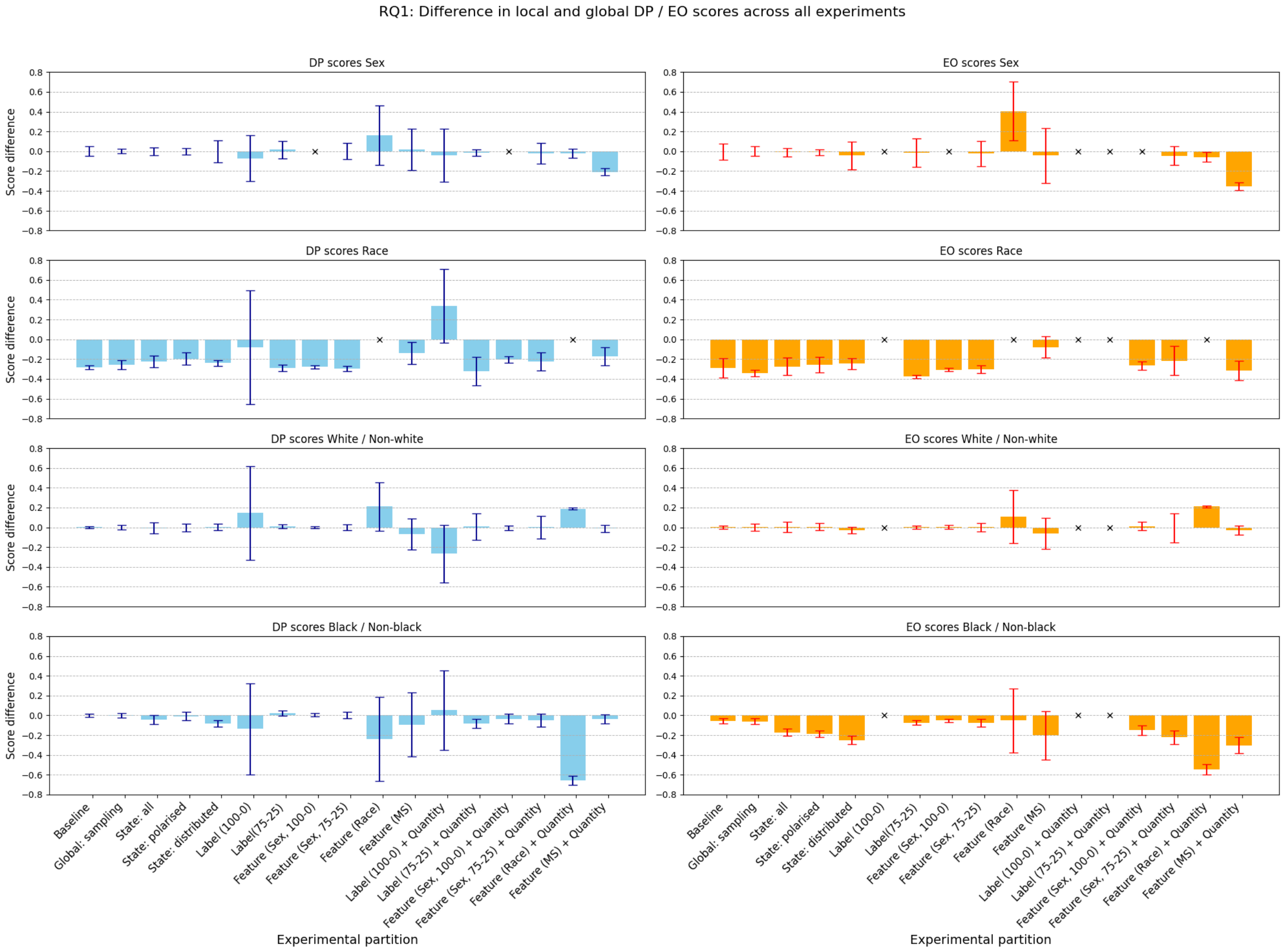


Figure 5.1: Comparison of demographic parity (DP, blue) and equalized odds (EO, orange) score differences across various experimental client partitions for different sensitive attributes categorizations. Scores are reported as the average difference between local and global pipeline across five runs with different random seeds. Each subplot represents the scores with respect to a specific attribute (Sex, Race, White/Non-white, Black/Non-black). Error bars represent the standard deviation across the five runs.

Table 5.2: Bias comparison between decentralized training with global bias assessment (global) and centralized training (central) for baseline data partition as the average over five random seeds, and the corresponding standard deviations.

	Central	**Global**	**Difference**
Accuracy	0.731 (±0.022)	0.735 (±0.001)	+0.004 (±0.022)
DP			
Sex	0.599 (±0.104)	0.668 (±0.034)	+0.069 (±0.109)
Race	0.351 (±0.110)	0.347 (±0.010)	-0.004 (±0.110)
White	0.692 (±0.065)	0.592 (±0.006)	-0.100 (±0.065)
Black	0.690 (±0.087)	0.679 (±0.014)	-0.011 (±0.088)
EO			
Sex	0.581 (±0.110)	0.711 (±0.060)	+0.130 (±0.125)
Race	0.324 (±0.078)	0.290 (±0.096)	-0.034 (±0.124)
White	0.672 (±0.080)	0.523 (±0.013)	-0.149 (±0.081)
Black	0.793 (±0.066)	0.779 (±0.013)	-0.014 (±0.067)

5.2.2 Heterogeneity

To evaluate the impact of heterogeneity, we compared the bias from the decentralized training of a global model to the bias from a centrally trained model across different heterogeneous client partitions. The absolute differences between the pipelines are reported for all experimental partitions in Figure 5.2. The average differences over five runs with different random seeds are presented along with their corresponding standard deviations.

The results from the experimental partitions do not show a consistent pattern similar to the baseline. Different experimental partitions demonstrate varying levels of improvement and deterioration of fairness within the federated learning pipeline compared to the central pipeline. Some partitions show significant standard deviations.

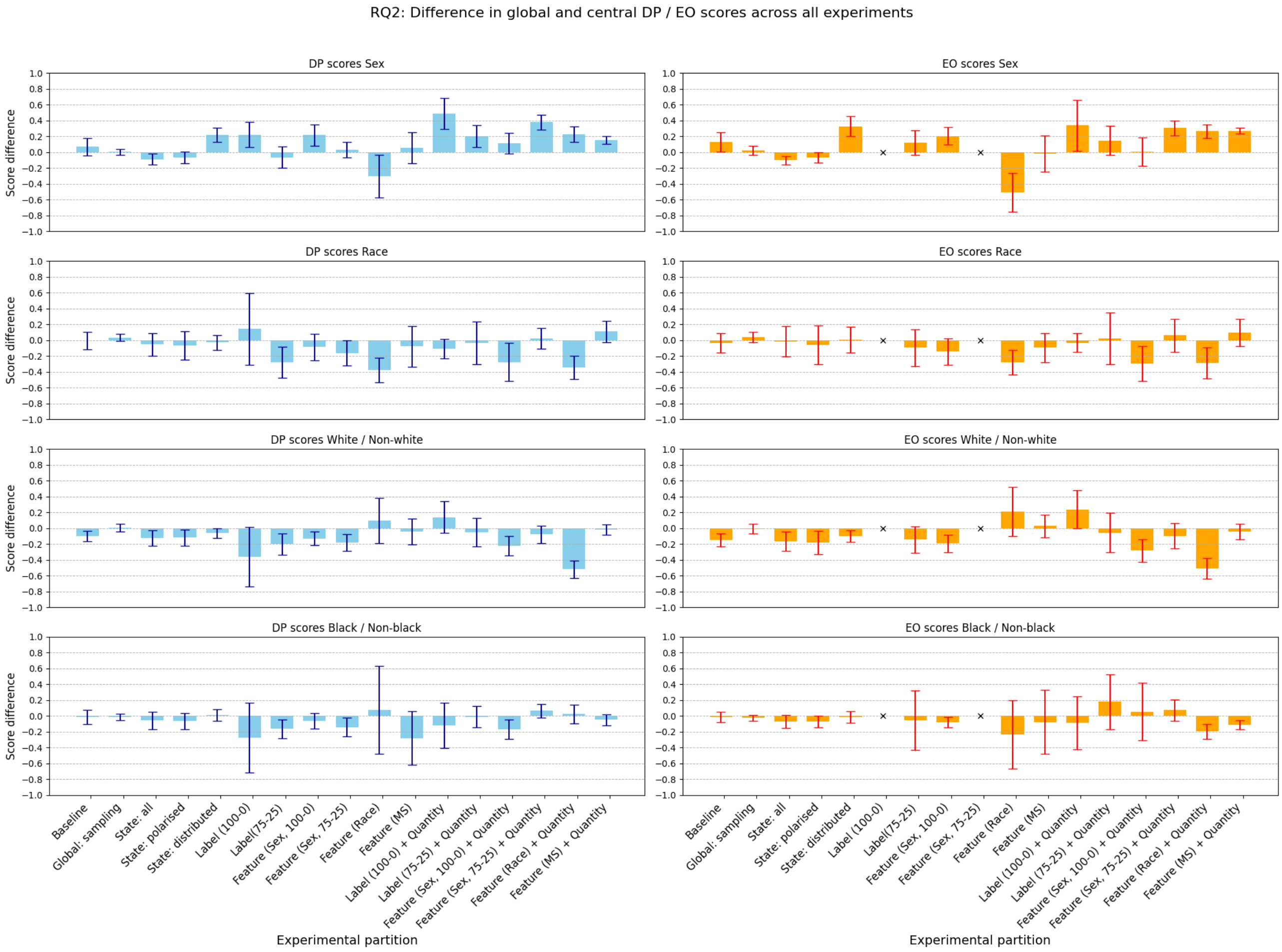


Figure 5.2: Comparison of demographic parity (DP, blue) and equalized odds (EO, orange) score differences across various experimental client partitions for different sensitive attributes categorizations. Scores are reported as the average difference between global and central pipelines across five runs with different random seeds. Each subplot represents the scores with respect to a specific attribute (Sex, Race, White/Non-white, Black/Non-black). Error bars represent the standard deviation across the five runs.

Chapter 6

Discussion

The aim of this research was two-fold: first, to evaluate the performance of the aggregated local bias assessment method, and second, to assess the emergence of bias in federated learning compared to centralized learning. The experiments for both objectives revealed discrepancies between the pipelines.

Due to the federated learning and data partitioning setup, comparing these results to previous research is challenging. However, with additional experiments on the 2018 ACS Income dataset, it is possible to compare the accuracy of the original ACS PUMS paper (Ding et al., 2021) with the accuracy of our centralized learning pipeline. From this, it can be concluded that the found accuracy of 72.6% (±0.06) is relatively similar to the one reported by Ding et al. (2021) (77.1%). Our central pipeline does not allow for a comparison of bias assessment results with the original paper, due to differences in their used classifier and race labels for their bias assessment. Comparing the results of both federated learning pipelines to existing work is also complicated, due to the limited literature on bias assessment in federated learning.

For the main experiments on the 2022 version of the ACS Income dataset, the test accuracies across the three pipelines are similar. Generally, across all experiments, the accuracy measured by the aggregated local method and the global method is identical. In most cases, these accuracies also match the performance of the centrally trained model, except for a few experimental partitions with extreme heterogeneities. These results are reported in Figure C.1 in the appendix.

6.1 Evaluation of Aggregated Local Bias Assessment

For the first research question, it is expected that there is a difference between the aggregated local bias assessment and the global bias assessment. It is expected that this difference is even more exacerbated for heterogeneous client partitions. Generally, the differences between the two assessment methods are relatively small, except for some sensitive attribute measurements that will be discussed in more detail below. Overall, both the global and aggregated local assessment measure more bias (lower bias scores) with respect to Race, compared to the Sex, White/Non-white and Black/Non-black attributes. This can be seen in Table 5.1 for the baseline and in Appendix C for the other experiments. The more prevalent bias with respect to Race can possibly be explained by the non-equal class sizes for this sensitive attribute (see Figure 4.1), that possibly lead to over- and underrepresentation of some race groups, thereby introducing bias. This observation can potentially indicate a historical bias.

6.1.1 Baseline

The baseline data partition compares the aggregated local and global bias assessment for a client partition that aimed for complete homogeneity in quantity, feature and label. The results of this partition, as presented in Table 5.1, show low standard deviations, indicating stability over different random seeds. The aggregated local bias assessment of this partition measured significantly more race bias (bias score reduction of 30%) compared to the global bias assessment. However, the Race bias score measured by the global assessment is already almost twice as low compared to the bias scores for the other sensitive attributes, which shows that the federated model is more biased with respect to this attribute. The discrepancy in the DP and EO metrics regarding Race can possibly be explained by the distribution of demographic groups for this sensitive attribute. From Figure 4.1, it can be seen that some race classes (e.g. *white* and *black*) are much larger than others (e.g. *Alaska Native* and *American Indian*). These less-frequent race classes will be underrepresented in some client datasets, which can lead to more bias. This can potentially be aggregated when using smaller client datasets for measuring bias, compared to using the global dataset. For the more equally distributed sex-classes, there is less bias and a smaller difference between the pipelines.

The local EO score for the Black/Non-black categorization is approximately 5.5% lower compared to its globally measured equivalent. This trend is not observed for the DP score. This indicates that the number of positive outcomes (i.e. the DP score, that represents how often an income of >$50k is predicted) between *black* and *non-black* is the same in the local and global pipeline, but that the performance with respect to these outcomes (i.e. the EO score, that is represented as the false positive rate or true positive rate) differs between the pipelines. This tendency can again be explained by a relatively disproportionate distribution between the *black* and *non-black* class (see Figure 4.1), which in this case specifically resulted in a disparity in false positive rate and true positive rate across the two classes. Furthermore, for the White/Non-white categorization, the differences in the DP score and EO score between the pipelines are more similar (only differ max. 0.2%). This categorization is more equally distributed, which supports the idea that the differences between the aggregated local and global pipelines can be explained by the disproportionate distribution of sensitive attribute classes.

6.1.2 Heterogeneity

Even though it was expected that the absolute difference between the global and aggregated local bias assessment would increase with the introduction of data heterogeneity, this trend is not clearly visible for the heterogeneous client partitions.

The client partitions with quantity heterogeneity show results very similar to the baseline. Figure 5.1 shows that the negative difference between the local and global bias assessments across all metrics is only slightly affected by partitions with quantity heterogeneity (experiments 2, 3, 4, and 5) compared to the baseline, but the differences between the partitions is small (max. 11%, as calculated using the tables in Appendix C). Again, the discrepancies are more pronounced for the DP and EO scores with respect to Race, and the EO score of the Black/Non-black categorization. This suggests that data heterogeneity in quantity, and how this is introduced (state-based or global-based), has minimal influence on the discrepancy between the global and aggregated local assessment.

The federated learning models that used full label heterogeneity (100-0) showed missing local EO scores and very high standard deviations, as shown in Figure 5.1. An analysis of the accuracy of the trained FL models for this partition, also revealed random accuracy scores, as can be seen in Figure C.1. With full label heterogeneity, each client contains only one label class, causing local training to overfit to that class, which can cause the low accuracy and unstable outcomes. This caused the absolute differences between the pipelines to vary significantly between runs, which makes it challenging to draw conclusions. In contrast, label heterogeneity with a 75-25 split per client provides much more stable results, as can be seen from the standard deviations in Table C.11. By introducing two label classes in each client dataset, this partition allows for local EO scores and less overfitting during local training. This data partition shows a pattern that is similar to the baseline, where the aggregated local assessment of the DP and EO scores for Race, and the EO score for Black/Non-black measure more bias compared to the global assessment.

The federated learning models that used data heterogeneity in feature distribution show a similar pattern, as seen in Figure 5.1. However, the data partition `Feature (Race)` and `Feature (MS)`[1] yield relatively unstable results, with relatively high standard deviations. This can also be verified by taking a closer look at the individual metric scores (and their standard deviations) as shown in Table C.17 and Table C.19. This can possibly be explained by the high number of data samples disregarded for feature heterogeneity due to a cut-off in these groups (see Table B.1). The partitions based on Race and Marital Status will consist of smaller local datasets (and thus a smaller global dataset), which could explain the more unstable results. This is also reflected in the accuracies of these partitions, which are slightly lower compared to most other partitions (see Figure C.1). A closer look at the standard deviations of the individual pipelines for these partitions reveals that the local pipeline is slightly more stable than the global pipeline, although both show significant standard deviations. This makes it hard to draw any concrete conclusions from these data partitions.

A combination of heterogeneities (i.e. heterogeneity in quantity and label/feature), does not seem to significantly influence the measured differences, as shown in Figure 5.1, supporting the results from the individual quantity heterogeneities that revealed no clear disparity compared to the baseline.

For the experimental partitions that have heterogeneity in `Feature (Race)`, `Label + Quantity`, `Feature (Race) + Quantity`, `Feature (Sex) + Quantity`, the DP and EO scores with respect to Black/Non-black and White/Non-white appear to be complementary; in cases where there is a positive difference for one categorization, there is often a negative difference for the other. This means that if the aggregated local assessment measures more bias with respect to White/Non-white compared to the global assessment, it is the other way around for Black/Non-black. It is expected that the bias scores for White/Non-white are generally lower (thus less fair) compared to the scores for the Black/Non-black categorization, since the *white* and *non-white* classes have a clearer distinction in unprotected and protected data samples compared to *black* and *non-black*. However, the results show that this is not always the case, and that this can also differ between the global and aggregated local assessment, as seen in the tables for data partition 6, 12 and 16 as presented in Appendix C. These findings indicate that some client partitions are more fair with respect to the Black/Non-black categorization than to the White/Non-white categorization (e.g. client partition 10, 11, 12), but that these nuances are not always measured with the aggregated local bias assessment.

6.1.3 Overview

In conclusion, for some sensitive attributes, the global bias assessment and aggregated local bias assessment yield similar bias scores. However, this consistency does not extend to both equalized odds and demographic parity measurements across all sensitive attribute categorizations. In the case of significant differences, the aggregated local bias assessment measures more bias (i.e. lower bias score) compared to the global assessment. This indicates that the use of the aggregated local bias assessment can potentially introduce an evaluation bias. Given that the sensitive attribute classes of Race and Black/Non-black are unevenly distributed, and

[1] A heterogeneity in `Marital Status` is in the figure abbreviated with `MS`.

that there are larger differences between the two pipelines for these categorizations, we can hypothesize that the disparity between aggregated local and global bias assessments increases in the case of unequal sensitive class distributions. This observation does not indicate the influence of heterogeneity, as the unequal distribution is consistent across all clients and thus does not represent client heterogeneity. Contrary to expectations, the introduction of client heterogeneity did not result in larger differences.

6.2 Bias in Federated Learning vs. Centralized Learning

For the second research question, it is expected that the difference between the global pipeline and the central pipeline will be prominent, showing more bias in the federated learning pipeline with global bias assessment (i.e. a negative difference). Furthermore, it is expected that this bias will be exacerbated with the introduction of data heterogeneity. Although in general, the differences between the two pipelines seem less consistent compared to the differences from the first research objective, here again, both the central and global pipeline measure more bias (i.e. lower bias scores) with respect to Race, compared to the Sex, White/Non-white and Black/Non-black attributes. This can be seen in Table 5.2 for the baseline and in Appendix C for the other experiments. This trend can again be explained by the less-equally distributed race classes, that potentially indicates a historical bias.

6.2.1 Baseline

The baseline data partition compares the bias from a global FL model and a centrally trained model for a client partition that aimed for complete homogeneity in quantity, feature and label. As seen in Table 5.2, the accuracy scores of the central and global pipelines are very similar, with only a slightly higher accuracy of 0.4% for the global FL model. Both the DP and EO metrics show that the global pipeline is more fair with respect to Sex, an increase of 6.9% and 13% respectively, compared to the central pipeline. For the other sensitive attributes, the metrics show slightly more bias (difference of max 15%) in the global FL model. However, these differences are small, and Table 5.2 shows significant standard deviations for these results. Nonetheless, these results indicate that federated learning can introduce bias, which is in line with the conclusion from Chang and Shokri (2023).

6.2.2 Heterogeneity

In Figure 5.2 it can be seen that the DP and EO scores with respect to the attribute Sex show a positive difference for most client partitions, indicating that the central pipeline is more biased with respect to Sex, compared to the global FL pipeline. For the other sensitive attributes, more bias is measured in the global model (i.e. a negative difference is shown), although this is not consistent across all experimental partitions. The differences in bias between the global and central pipeline fluctuate between the different experimental partitions. For most experimental partitions there is a large standard deviation measured between the different seeds. This variability means that the observed tendencies from these results are not very reliable and thus less generalizable to other experimental runs.

The experimental partitions that have quantity heterogeneity do not show a clear indication of increased bias in either the global pipeline or the central pipeline. Generally, the differences between the pipelines seem relatively small, and comparable to the baseline. This indicates that quantity heterogeneity does not have a significant influence on bias in federated learning.

Figure 5.2 shows very large standard deviations over the pipeline differences for the experimental setups with heterogeneity in `Label (100-0)` and `Label (100-0) + Quantity`. The instability in data partitions using full label heterogeneity can be attributed to high instability in the federated learning pipeline. This is evident from Table C.10, where the standard deviation for the global FL pipeline is very prominent and larger than the standard deviation of the central pipeline. The instability of the FL model can be assigned to each client containing only one label class, which leads to local overfitting to that class.

The results for data partitions with feature heterogeneity do not show a clear indication of increased bias in the global FL model. Some of these partitions, i.e. `Feature (Race)` and `Feature (Marital Status)`[2], show significant standard deviations on their resulting differences. The high standard deviations for these partitions can be explained by the low number of data samples used for these partitions. Partitions with heterogeneity in Race and Marital Status have much smaller datasets due to the cut-off of data samples to obtain the partitions with heterogeneity in feature, but not in quantity (see Figure B.1). These smaller partitions lead to less stable results and thus higher standard deviations for both the global and central pipeline, as can be seen in more detail in Table C.18 and Table C.20.

6.2.3 Overview

In conclusion, the obtained differences between the global and central pipeline do not show a consistent and significant similarity or discrepancy between the global and central pipelines across different runs. These results indicate that federated learning can perpetuate more biases, but that this is not necessarily the case in all scenarios (i.e. for all runs and sensitive attributes). The detection of additional bias by the global pipeline in some scenarios indicates that federated learning carries a risk of introducing algorithmic bias. Additionally, the introduction of data heterogeneity does not appear to influence the discrepancy between federated learning and centralized learning.

[2] A heterogeneity in `Marital Status` is in the figure abbreviated with `MS`.

Chapter 7

Conclusions and Future Work

Motivated by the current challenges of fairness research in federated learning, this study aimed to address two main research gaps in current literature. The first goal was to evaluate the accuracy of the proposed aggregated local bias assessment methodology, that is used to measure bias in federated learning frameworks without compromising privacy, by answering the research question: *How accurate is the aggregation bias assessment method in detecting bias that arises in a federated learning framework?* Secondly, this study tried to assess whether a federated learning framework introduces more or less bias compared to a central learning framework, by answering the question: *What bias arises during the federated learning process compared to a centralized learning process?* This research thereby aims to bridge the gap between private federated learning and fair centralized learning to advance towards safe and responsible AI. It was hypothesized that the results for both objectives would be influenced by the presence of data heterogeneity within client partitions.

The results for the first research objective showed negative discrepancies between the aggregated local and global bias assessment method for some sensitive attributes, meaning that the aggregated local bias assessment measures more bias compared to the global assessment. This discrepancy seems most prevalent for sensitive attributes that have unequal sensitive class distributions. The results for the second research question failed to show a consistent significant tendency towards either the central learning pipeline or the global federated learning pipeline. However, in some scenarios the global FL pipeline did measure more bias, indicating that federated learning can introduce bias with its decentralized training nature.

As opposed to what was hypothesized, the introduction of heterogeneity did not yield consistently different results for both research objectives, but often lead to more unstable and less reliable results. With the choice of a real-world dataset with a natural partition, the introduction of different heterogeneous partitions becomes a complex task, where it is not possible to guarantee complete homogeneity and exact heterogeneity across clients. When creating heterogeneity for one component (e.g. a feature) manually, heterogeneity can be automatically introduced in a different component (e.g. label) when both components show a strong correlation in the original dataset. The current results were not able to measure a consistent tendency towards specific types of heterogeneity, but the noise of potentially undesired heterogeneity from the automatically varying component, can possibly skew these results. Hence, investigating the correlation between the heterogeneous component and the dynamically varying component, and its impact on bias assessment results, is a relevant topic for future research. Furthermore, the extension of this research to other datasets lacking natural partitions, where partitions can be artificially created, is also proposed as future research.

Another limitation of the current dataset is the sensitivity of the used income threshold. Previous literature has shown that using an income threshold of $50k for fairness research on income prediction can result in an inadequate and misleading representation of the broader context (Ding et al., 2021). They found that in many cases, a threshold of $50k misrepresents the broader picture that other thresholds captured. It would be interesting to see whether this sensitivity also translates to the federated learning frameworks. Future research can look at the influence of the label threshold on the different pipeline.

Besides the influence of data heterogeneity on bias in federated learning, earlier research proposed client selection as a configuring factor in the emergence of bias within federated learning. This research has not addressed this potential issue. Future research could investigate how various client selection methods affect the emergence and assessment of bias in FL models. Additionally, comparing the results of the FedAvg aggregation algorithm with those obtained using other aggregation algorithms would be valuable to determine if there are significant differences in bias. Furthermore, the generalization of the current approaches to different fairness metrics, fairness types (i.e. individual or subgroup fairness), training applications and datasets is left for future research.

To conclude, this research has established that the *aggregated local bias assessment* method does not consistently measure the same bias scores as the accurately established *global bias assessment*. The next step would be to probe where these differences are coming from and in what scenario's they occur. Hereby, the goal is to obtain a bias assessment methodology that allows for an accurate, privacy-preserving and global measurement of bias within federated learning. Furthermore, this research has shown that federated learning can introduce more bias compared to central learning, although this is not the case for every metric, sensitive attribute and client partition. Moving forward, a promising direction is to identify what part of federated learning contributes to this bias, such that effective bias mitigation methods can be established for federated learning. This thesis has taken promising initial steps towards comprehensively understanding bias assessment in federated learning. It is hoped that these findings will inspire further exploration in this area, thereby contributing to the development of fair, privately-trained AI models.

Bibliography

Annie Abay, Yi Zhou, Nathalie Baracaldo, Shashank Rajamoni, Ebube Chuba, and Heiko Ludwig. Mitigating bias in federated learning. *arXiv preprint arXiv:2012.02447*, 2020.

Annie Abay, Ebube Chuba, Yi Zhou, Nathalie Baracaldo, and Heiko Ludwig. Addressing unique fairness obstacles within federated learning. *AAAI RDAI-2021*, 2021.

Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. *Advances in neural information processing systems*, 32, 2019.

Adrien Banse, Jan Kreischer, et al. Federated learning with differential privacy. *arXiv preprint arXiv:2402.02230*, 2024.

Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*, 2017.

Christina Hastings Blow, Lijun Qian, Camille Gibson, Pamela Obiomon, and Xishuang Dong. Comprehensive validation on reweighting samples for bias mitigation via aif360. *Applied Sciences*, 14(9):3826, 2024.

William Briguglio, Parisa Moghaddam, Waleed A Yousef, Issa Traoré, and Mohammad Mamun. Machine learning in precision medicine to preserve privacy via encryption. *Pattern Recognition Letters*, 151:148–154, 2021.

Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. Understanding the origins of bias in word embeddings. In *International conference on machine learning*, pages 803–811. PMLR, 2019.

David Byrd and Antigoni Polychroniadou. Differentially private secure multi-party computation for federated learning in financial applications. In *Proceedings of the First ACM International Conference on AI in Finance*, pages 1–9, 2020.

Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. *Advances in neural information processing systems*, 30, 2017.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX security symposium (USENIX security 19)*, pages 267–284, 2019.

Jillian Carmody, Samir Shringarpure, and Gerhard Van de Venter. Ai and privacy concerns: a smart meter case study. *Journal of Information, Communication and Ethics in Society*, 19 (4):492–505, 2021.

Dario Catalano, Ronald Cramer, Giovanni Di Crescenzo, Ivan Darmgård, David Pointcheval, Tsuyoshi Takagi, Ronald Cramer, and Ivan Damgård. Multiparty computation, an introduction. *Contemporary cryptology*, pages 41–87, 2005.

Zheng Chai, Ahsan Ali, Syed Zawad, Stacey Truex, Ali Anwar, Nathalie Baracaldo, Yi Zhou, Heiko Ludwig, Feng Yan, and Yue Cheng. Tifl: A tier-based federated learning system. In *Proceedings of the 29th international symposium on high-performance parallel and distributed computing*, pages 125–136, 2020.

Hongyan Chang and Reza Shokri. Bias propagation in federated learning. *arXiv preprint arXiv:2309.02160*, 2023.

Yae Jee Cho, Jianyu Wang, and Gauri Joshi. Client selection in federated learning: Convergence analysis and power-of-choice selection strategies. *arXiv preprint arXiv:2010.01243*, 2020.

Arka Rai Choudhuri, Matthew Green, Abhishek Jain, Gabriel Kaptchuk, and Ian Miers. Fairness in an unfair world: Fair multiparty computation from public bulletin boards. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 719–728, 2017.

Alexandra Chouldechova and Aaron Roth. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*, 2018.

Rachel Cummings, Varun Gupta, Dhamma Kimpara, and Jamie Morgenstern. On the compatibility of privacy and fairness. In *Adjunct publication of the 27th conference on user modeling, adaptation and personalization*, pages 309–315, 2019.

Xolani Dastile, Turgay Celik, and Moshe Potsane. Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing*, 91:106263, 2020.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in neural information processing systems*, 34:6478–6490, 2021.

Wei Du, Depeng Xu, Xintao Wu, and Hanghang Tong. Fairness-aware agnostic federated learning. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pages 181–189. SIAM, 2021.

Cynthia Dwork. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, pages 1–19. Springer, 2008.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pages 265–284. Springer, 2006.

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.

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.

Evanthia Faliagka, Kostas Ramantas, Athanasios Tsakalidis, and Giannis Tzimas. Application of machine learning algorithms to an online recruitment system. In *Proc. International Conference on Internet and Web Applications and Services*, pages 215–220, 2012.

Haokun Fang and Quan Qian. Privacy preserving machine learning with homomorphic encryption and federated learning. *Future Internet*, 13(4):94, 2021.

Lynda Ferraguig, Yasmine Djebrouni, Sara Bouchenak, and Vania Marangozova. Survey of bias mitigation in federated learning. In *Conférence francophone d'informatique en Parallélisme, Architecture et Système*, 2021.

Emilio Ferrara. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *arXiv preprint arXiv:2304.07683*, 2023.

Ferdinando Fioretto, Cuong Tran, Pascal Van Hentenryck, and Keyu Zhu. Differential privacy and fairness in decisions and learning tasks: A survey. *arXiv preprint arXiv:2202.08187*, 2022.

Simson L Garfinkel, John M Abowd, and Sarah Powazek. Issues encountered deploying differential privacy. In *Proceedings of the 2018 Workshop on Privacy in the Electronic Society*, pages 133–137, 2018.

Yuhao Gu, Yuebin Bai, and Shubin Xu. Cs-mia: Membership inference attack based on prediction confidence series in federated learning. *Journal of Information Security and Applications*, 67:103201, 2022.

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

Jianping He and Lin Cai. Differential private noise adding mechanism: Basic conditions and its application. In *2017 American Control Conference (ACC)*, pages 1673–1678. IEEE, 2017.

Senerath Mudalige Don Alexis Chinthaka Jayatilake, Gamage Upeksha Ganegoda, et al. Involvement of machine learning tools in healthcare decision making. *Journal of healthcare engineering*, 2021, 2021.

Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of ai ethics guidelines. *Nature machine intelligence*, 1(9):389–399, 2019.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2):1–210, 2021.

Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1):1–33, 2012.

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23*, pages 35–50. Springer, 2012.

Yigitcan Kaya and Tudor Dumitras. When does data augmentation help with membership inference attacks? In *International conference on machine learning*, pages 5345–5355. PMLR, 2021.

Akbar Khanan, Salwani Abdullah, Abdul Hakim HM Mohamed, Amjad Mehmood, and Khairul Akram Zainol Ariffin. Big data security and privacy concerns: a review. In *Smart Technologies and Innovation for a Sustainable Future: Proceedings of the 1st American University in the Emirates International Research Conference—Dubai, UAE 2017*, pages 55–61. Springer, 2019.

Ronny Kohavi and Barry Becker. Adult data set. *UCI machine learning repository*, 5:2093, 1996.

Emmanouil Krasanakis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris. Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In *Proceedings of the 2018 world wide web conference*, pages 853–862, 2018.

Li Li, Yuxi Fan, Mike Tse, and Kuo-Yi Lin. A review of applications in federated learning. *Computers & Industrial Engineering*, 149:106854, 2020.

Pengfei Li, Yunfeng Zhao, Liandong Chen, Kai Cheng, Chuyue Xie, Xiaofei Wang, and Qinghua Hu. Uncertainty measured active client selection for federated learning in smart grid. In *2022 IEEE International Conference on Smart Internet of Things (SmartIoT)*, pages 148–153. IEEE, 2022.

Shuchang Liu, Yingqiang Ge, Shuyuan Xu, Yongfeng Zhang, and Amelie Marian. Fairness-aware federated matrix factorization. In *Proceedings of the 16th ACM conference on recommender systems*, pages 168–178, 2022.

Pranay Lohia. Priority-based post-processing bias mitigation for individual and group fairness. *arXiv preprint arXiv:2102.00417*, 2021.

Pranay K Lohia, Karthikeyan Natesan Ramamurthy, Manish Bhide, Diptikalyan Saha, Kush R Varshney, and Ruchir Puri. Bias mitigation post-processing for individual and group fairness. In *Icassp 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 2847–2851. IEEE, 2019.

Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.

Priyanka Mary Mammen. Federated learning: Opportunities and challenges. *arXiv preprint arXiv:2101.05428*, 2021.

Don R McCreary. Cambridge academic content dictionary. *Dictionaries: Journal of the Dictionary Society of North America*, 30(1):151–155, 2009.

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.

Ninareh Mehrabi, Umang Gupta, Fred Morstatter, Greg Ver Steeg, and Aram Galstyan. Attributing fair decisions with attention interventions. *arXiv preprint arXiv:2109.03952*, 2021a.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6): 1–35, 2021b.

Matias Mendieta, Taojiannan Yang, Pu Wang, Minwoo Lee, Zhengming Ding, and Chen Chen. Local learning matters: Rethinking data heterogeneity in federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8397–8406, 2022.

Adrian Nilsson, Simon Smith, Gregor Ulm, Emil Gustavsson, and Mats Jirstrand. A performance evaluation of federated learning algorithms. In *Proceedings of the second workshop on distributed infrastructures for deep learning*, pages 1–8, 2018.

Takayuki Nishio and Ryo Yonetani. Client selection for federated learning with heterogeneous resources in mobile edge. In *ICC 2019-2019 IEEE international conference on communications (ICC)*, pages 1–7. IEEE, 2019.

David Pujol, Ryan McKenna, Satya Kuppam, Michael Hay, Ashwin Machanavajjhala, and Gerome Miklau. Fair decision making using privacy-protected data. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 189–199, 2020.

Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh. Fairbatch: Batch selection for model fairness. *arXiv preprint arXiv:2012.01696*, 2020.

Stuart J Russell and Peter Norvig. *Artificial intelligence: a modern approach*. Pearson, 2016.

Sina Shaham, Arash Hajisafi, Minh K Quan, Dinh C Nguyen, Bhaskar Krishnamachari, Charith Peris, Gabriel Ghinita, Cyrus Shahabi, and Pubudu N Pathirana. Holistic survey of privacy and fairness in machine learning. *arXiv preprint arXiv:2307.15838*, 2023.

Yuxin Shi, Han Yu, and Cyril Leung. Towards fairness-aware federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

Reid G Smith and Joshua Eckroth. Building ai applications: Yesterday, today, and tomorrow. *Ai Magazine*, 38(1):6–22, 2017.

Cuong Tran, Ferdinando Fioretto, Pascal Van Hentenryck, and Zhiyan Yao. Decision making with differential privacy under a fairness lens. In *IJCAI*, pages 560–566, 2021.

Nguyen Truong, Kai Sun, Siyao Wang, Florian Guitton, and YiKe Guo. Privacy preservation in federated learning: An insightful survey from the gdpr perspective. *Computers & Security*, 110:102402, 2021.

Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the international workshop on software fairness*, pages 1–7, 2018.

Christina Wadsworth, Francesca Vera, and Chris Piech. Achieving fairness through adversarial learning: an application to recidivism prediction. *arXiv preprint arXiv:1807.00199*, 2018.

Angelina Wang, Alexander Liu, Ryan Zhang, Anat Kleiman, Leslie Kim, Dora Zhao, Iroha Shirai, Arvind Narayanan, and Olga Russakovsky. Revise: A tool for measuring and mitigating bias in visual datasets. *International Journal of Computer Vision*, 130(7):1790–1810, 2022.

Yongkai Wu, Lu Zhang, and Xintao Wu. Fairness-aware classification: Criterion, convexity, and bounds. *arXiv preprint arXiv:1809.04737*, 2018.

Yun Xie, Peng Li, Chao Wu, and Qiuling Wu. Differential privacy stochastic gradient descent with adaptive privacy budget allocation. In *2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, pages 227–231. IEEE, 2021.

Bin Yu, Wenjie Mao, Yihan Lv, Chen Zhang, and Yu Xie. A survey on federated learning in data mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12 (1):e1443, 2022.

Sergey Zapechnikov. Secure multi-party computations for privacy-preserving machine learning. *Procedia Computer Science*, 213:523–527, 2022.

Yuchen Zeng, Hongxu Chen, and Kangwook Lee. Improving fairness via federated learning. *arXiv preprint arXiv:2110.15545*, 2021.

Daniel Yue Zhang, Ziyi Kou, and Dong 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.

Ying Zhao and Jinjun Chen. A survey on differential privacy for unstructured data content. *ACM Computing Surveys (CSUR)*, 54(10s):1–28, 2022.

Appendix A

Dataset

A.1 Disregarded Datasets

Table A.1 shows a list of all considered datasets and the reasons for why they were eventually disregarded.

Table A.1: All datasets that were originally considered but disregarded due to significant disadvantages.

Dataset	**Reason**
MIMIC-III (link)	Lengthy application process for ethical credentialing.
Heart Disease dataset (link)	Only sensitive information on *Sex*, no known information on biases within this dataset.
Heart Attack Risk Prediction Dataset (link)	Artificially created dataset that has no discernible pattern between features and labels.
Health and Demographics Dataset (link	No data on sensitive attributes.
Dutch Virtual Census 2001 (link)	Lacks clear documentation and easy access to data samples without manual pre-processing.

A.2 Data Extraction ACS PUMS

The ACS Income dataset is extracted from the American Community Survey (ACS) Public Use Microdata Sample (PUMS) from 2005 to 2022. This survey is sent to approximately 3.5 million US households from all 51 US states (including Puerto Rico). This allows for a dynamic fairness and discrimination analysis across space and time. This survey is sent out every year to collect information relating to ancestry, citizenship, education, employment, language proficiency, income, disability, and housing characteristics. The Public Use Microdata Sample includes answers regarding all subjects from a selected group of respondents. The participation for households in the ACS is mandatory under federal law. Responses are kept confidential and are subject to strict privacy regulations. The geographic details linked to each participant are restricted to prevent the re-identification of individuals. Additionally, several other disclosure control heuristics are implemented within the process. Individuals obtained a personal weight, which is used to re-weight the dataset such that it represents the general US population accurately. A comprehensive documentation is accessible on the US Census Bureau website [1].

[1] `https://www.census.gov/en.html`

The ACS Income dataset considers a classification objective in which the target feature is a binary indicator whether individuals earn below or above a $50k income threshold. There are nine different features considered for this objective. These features are shown in Table A.2.

Table A.2: Table of all used features, and their documentation.

Age	The age of the individual (ranges from 16 to 99).
Class of Worker	Individual's employment type.
Education	Educational level reached by individual.
Marital Status	Married, widowed, divorced, separated or never married.
Occupation	Current occupation of individual.
Place of Birth	Individual's place of birth, ranges from US states to most countries.
Worked hours	Usual hours worked per week for the past 12 months.
Sex	Male or female.
Race	Detailed race and ethnicity code as is practiced by US government.

A.3 Data Analysis ACS Income

This section shows a detailed overview of some specific data characteristics of the ACS Income dataset. Figure A.1 shows the number of individual data samples per US state. This image shows that there are significant size differences between states. The ACS Income dataset contains more individual data samples from more-populated states like Texas and California, compared to less-populated states like Wyoming and Alaska. This is also reflected in Figure A.2 which shows the ACS Income dataset size as the percentage of the population per state. This figure illustrates that the ACS Income dataset covers approximately 0.4% to 0.7% of the total population in each state. This means that the distribution of data samples per state in the ACS Income dataset is representable for the real world distribution of populations across all US states.

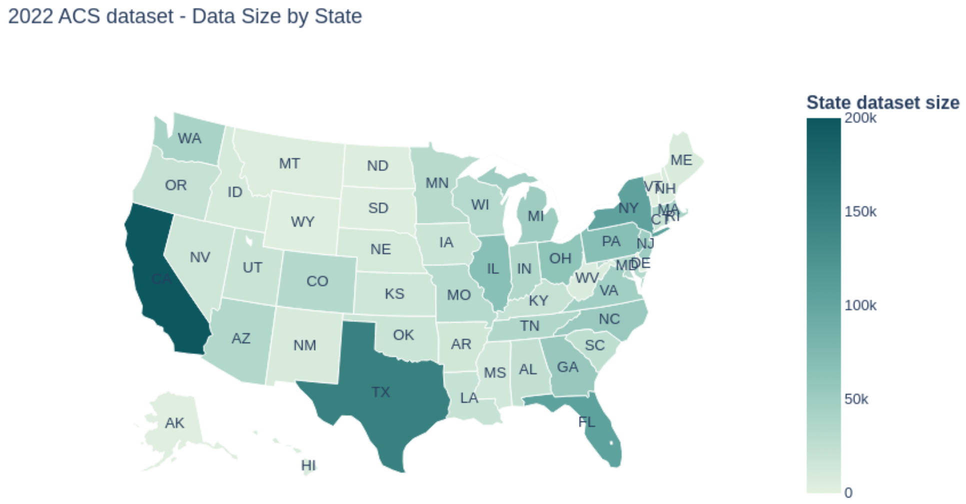


Figure A.1: The number of individual data samples in ACS Income dataset per US state.

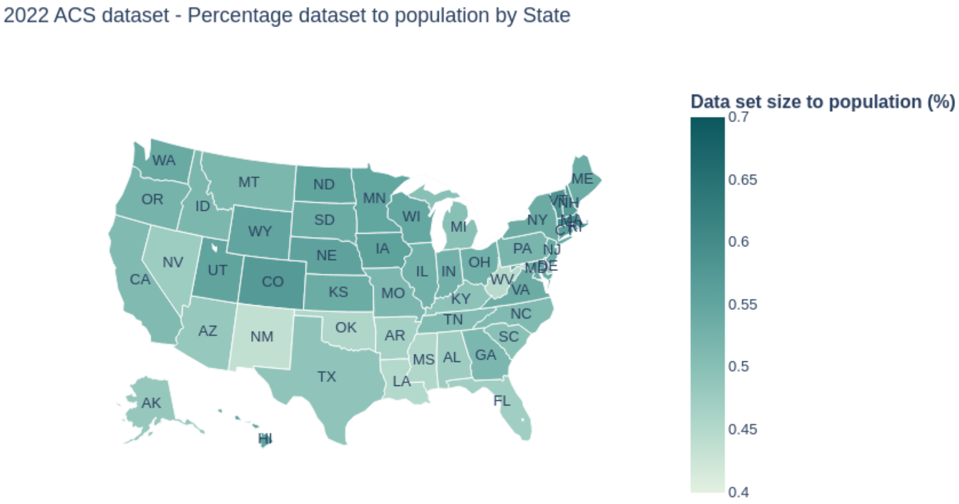


Figure A.2: The number of individual data samples from each state as a percentage of the entire population of that state.

Figure A.3 shows the distribution of number of individuals that earn above $50k, as a percentage of the entire number of data samples across for each state. This figure shows that most states have a balanced distribution of around 50% of the data samples having an income above $50k and around 50% of the data samples having an income below $50k.

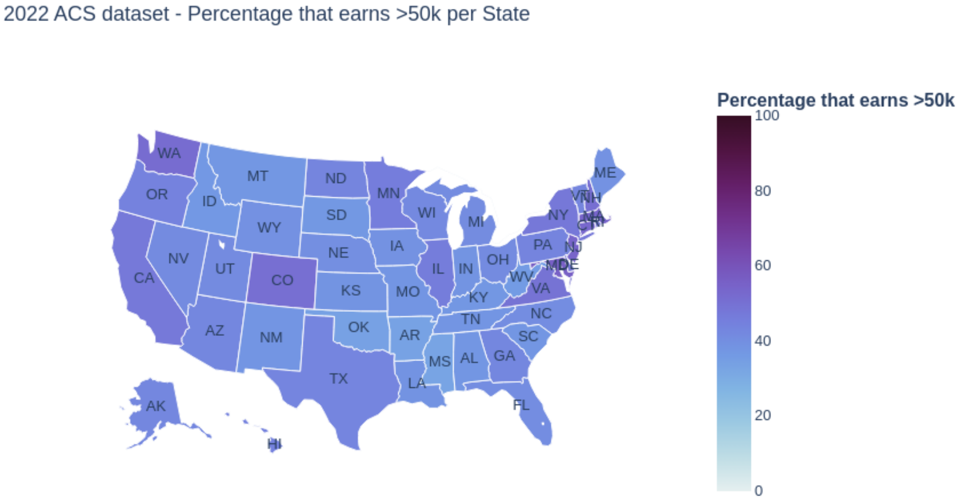


Figure A.3: Percentage of data samples that earns above $50k for each state.

Figure A.4 shows the distribution of female and male individuals across all states, and the distribution of males and females with an income above $50k. This visualization shows that the female and male population is roughly equally distributed, but that for each state, the percentage of male population with an income above $50k is higher compared to the female population with an income above $50k.

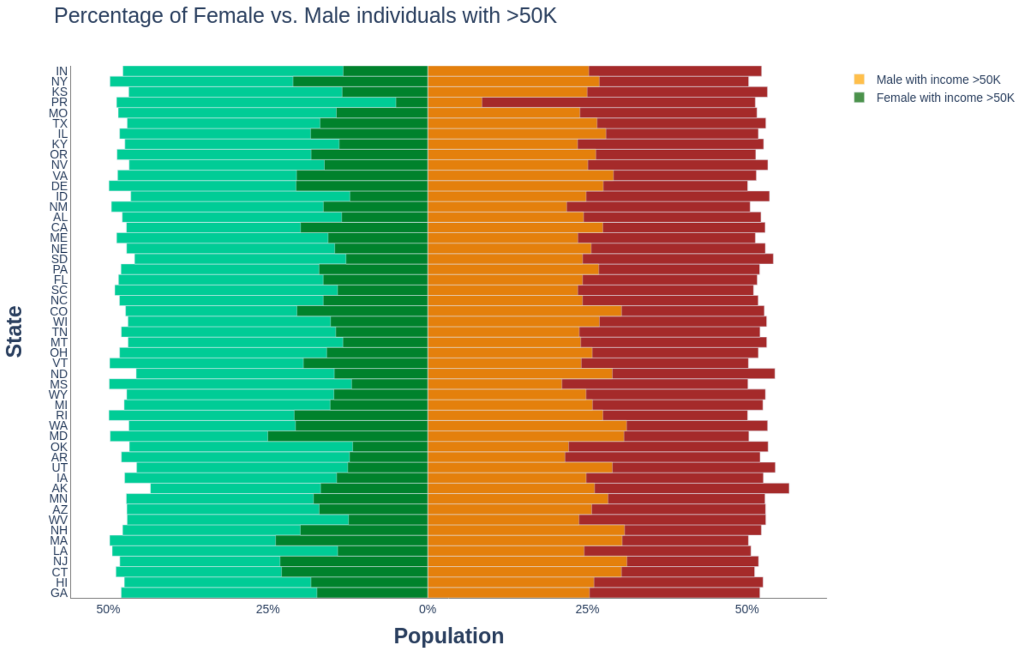


Figure A.4: Number of female and male individuals as percentage of total number of individuals per state, compared to the male and female class that have an income above $50k per US state.

Figure A.5 shows the distribution of *White* and *Non-white* individuals across all states, and the distribution of whites and non-whites with an income above $50k. This visualization shows that in most states the population of whites is bigger, but that this can differ per state. The percentage of whites with an income above $50k is also higher compared to the percentage of non-whites with an income above $50k. This unequal distribution of the white and non-white population, and their income, can lead to bias.

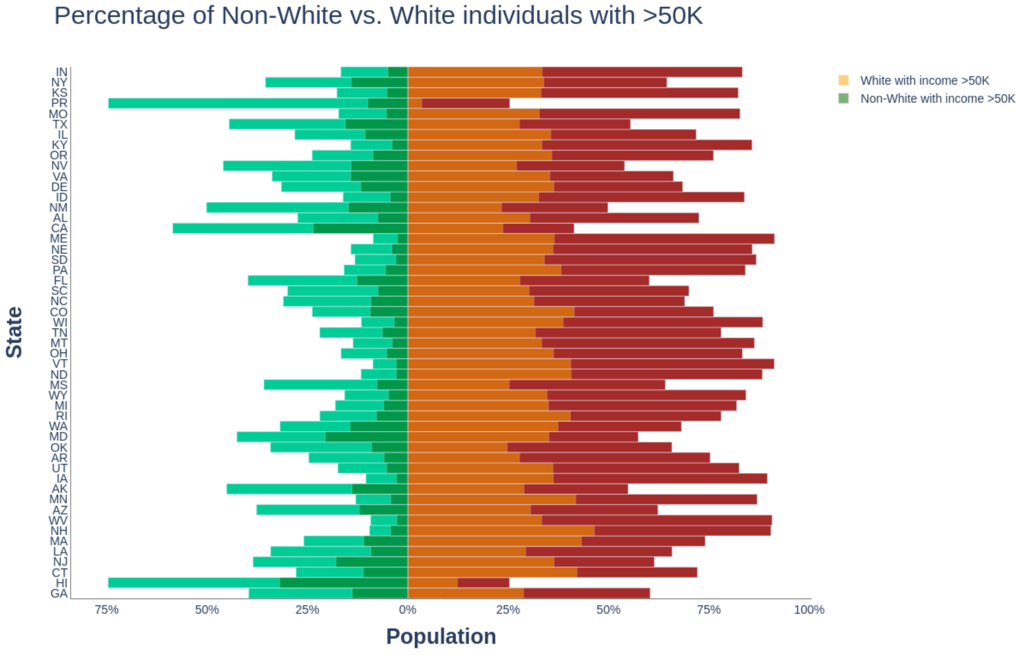


Figure A.5: Percentage of total non-whites and whites and percentage of non-whites/whites that earn above $50K (compared to total number of individuals for each state).

Appendix B

Setup Experimental partitions

Figure B.1 shows the global dataset sizes as obtained when exploiting the different experimental partitions. This image shows that the `Feature(Race)` and `Feature(Marital Status`[1]`)` partitions lead to a dataset that is only a small fraction of the original dataset size.

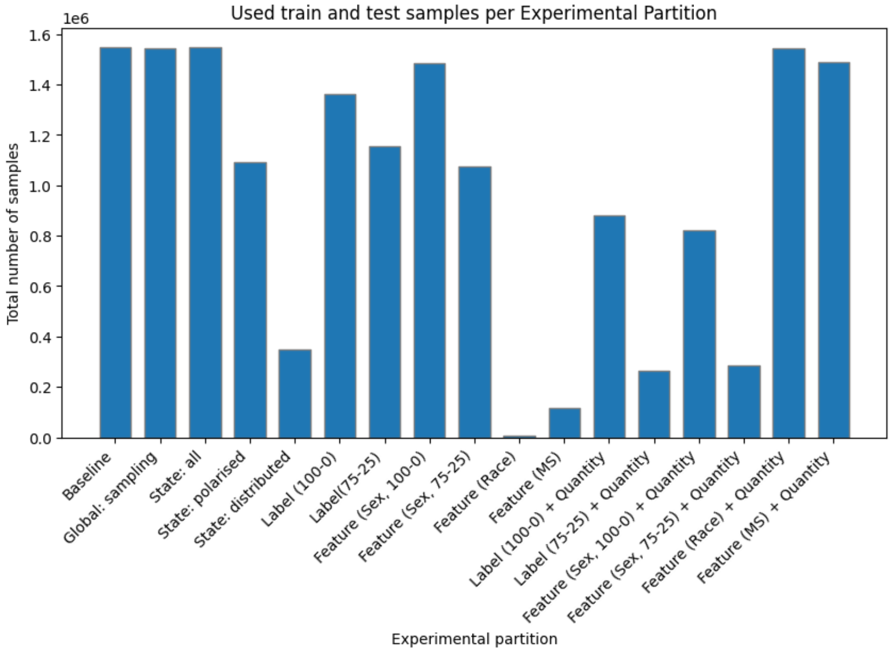


Figure B.1: Global dataset sizes across all experimental partitions.

Table B.1 shows how many data samples will be disregarding by a cut-off to create a client partition with feature heterogeneity and a uniform quantity. The features Sex and Marital Status have more equally distributed classes, such that these features have the least amount of data samples that will be disregarded by the cut-off, as shown in the table.

[1] In figure shown as`MS`.

Table B.1: Number of data samples that will be disregarded in the case of feature heterogeneity and uniform quantity, and how this relates to the total number of data samples (i.e. number of disregarded samples as percentage of the total number of data samples).

Feature	**Number of samples**	**Percentage of total number of samples**
Sex	71070	4.1%
Race	1711451	99.5%
Marital Status	1589519	92.4%
Education	1713146	99.6 %
Class of Worker	1684298	97.9 %
Occupation	1675718	97.4 %
Worked hours	1719659	99.9%

Appendix C

Results

C.1 Accuracies

Figure C.1 shows the accuracies from all three pipelines across all different experimental partitions. Results are shown as the average over five runs with different random seeds. Standard deviation over these averages are with an error bar.

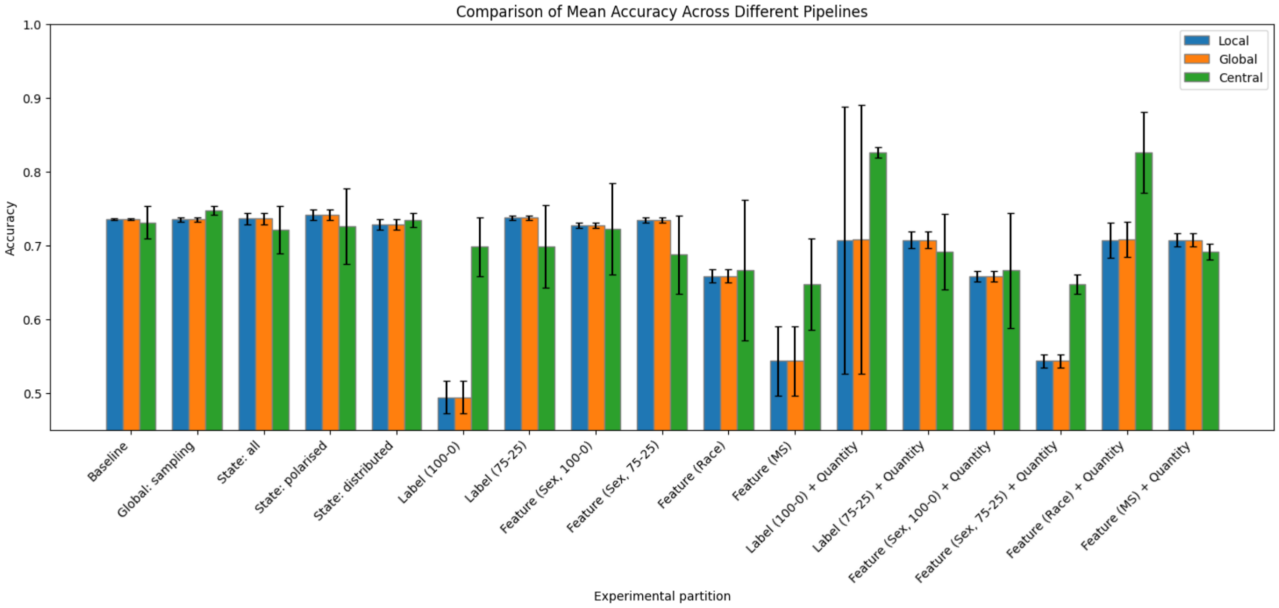


Figure C.1: Comparison of accuracy of the three pipelines across all experimental partitions. Average over five different random seeds.

C.2 Experimental Partitions

The following section will go over all detailed results for every experimental partition. For every partition, the metric scores are reported as the average of five runs with different random seeds. This is done for all three pipelines and their corresponding aboslute differences that relate to the two research objectives. Additionally, the standard deviations for all these are reported.

Partition 2: Heterogeneity in Quantity (Global-based: sampling datapoints)

Table C.1 and Table C.2 show the comparison between the global and local pipelines, and the comparison between the central and global pipelines, respectively, for the data partition with quantity introduced by sampling datapoints. The first comparison show a discrepancy between both pipelines for the EO and DP score of RACE, and for the EO score of BLACK/NON-BLACK. The second comparison show very small differences between the central and global pipeline for all metric scores.

Table C.1: Bias metric comparison between global bias assessment and aggregated local bias assessment for data partition with quantity heterogeneity introduced by sampling datapoints. Results are the average over five runs with different random seeds, and their corresponding standard deviation.

	Global	**Local**	**Difference**
Accuracy	0.735 (±0.003)	0.735 (±0.003)	0.000 (±0.004)
DP			
Sex	0.589 (±0.014)	0.590 (±0.015)	+0.001 (±0.021)
Race	0.348 (±0.027)	0.092 (±0.033)	-0.256 (±0.043)
White	0.603 (±0.015)	0.603 (±0.015)	0.000 (±0.021)
Black	0.665 (±0.017)	0.664 (±0.017)	-0.001 (±0.024)
EO			
Sex	0.579 (±0.035)	0.581 (±0.035)	+0.002 (±0.049)
Race	0.345 (±0.033)	0.003 (±0.003)	-0.342 (±0.033)
White	0.538 (±0.023)	0.539 (±0.023)	+0.001 (±0.033)
Black	0.772 (±0.017)	0.712 (±0.022)	-0.060 (±0.028)

Table C.2: Bias comparison between decentralized training and central training for data partition with quantity heterogeneity introduced by sampling datapoints. Results are the average over five runs with different random seeds, and their corresponding standard deviation.

	Central	**Global**	**Difference**
Accuracy	0.747 (±0.006)	0.735 (±0.003)	-0.012 (±0.007)
DP			
Sex	0.585 (±0.035)	0.589 (±0.014)	+0.004 (±0.038)
Race	0.314 (±0.032)	0.348 (±0.027)	+0.034 (±0.042)
White	0.599 (±0.044)	0.603 (±0.015)	+0.004 (±0.046)
Black	0.677 (±0.040)	0.665 (±0.017)	-0.012 (±0.043)
EO			
Sex	0.556 (±0.041)	0.579 (±0.035)	+0.023 (±0.054)
Race	0.305 (±0.054)	0.345 (±0.033)	+0.040 (±0.063)
White	0.546 (±0.053)	0.538 (±0.023)	-0.008 (±0.063)
Black	0.796 (±0.030)	0.772 (±0.017)	-0.024 (±0.034)

Partition 3: Heterogeneity in Quantity (State-based: using all states)

Table C.3 and Table C.4 show the comparison between the global and local pipelines, and the comparison between the central and global pipelines, respectively, for the data partition with quantity heterogeneity introduced by using all original states. The first comparison show a discrepancy between both pipelines for the EO and DP score of Race, and for the EO score of Black/Non-black. The second comparison show that for almost all metrics, the global FL pipelines is more fair compared to the central pipeline.

Table C.3: Bias metric comparison between global bias assessment and aggregated local bias assessment for data partition with quantity heterogeneity introduced by using all original states. Results are the average over five runs with different random seeds, and their corresponding standard deviation.

	Global	**Local**	**Difference**
Accuracy	0.739 (±0.008)	0.739 (±0.008)	0.000 (±0.076)
DP			
Sex	0.562 (±0.029)	0.561 (±0.029)	-0.001 (±0.041)
Race	0.330 (±0.047)	0.107 (±0.038)	-0.223 (±0.060)
White	0.583 (±0.038)	0.572 (±0.038)	-0.011 (±0.054)
Black	0.667 (±0.041)	0.661 (±0.021)	-0.006 (±0.046)
EO			
Sex	0.536 (±0.030)	0.527 (±0.028)	-0.009 (±0.041)
Race	0.286 (±0.086)	0.026 (±0.028)	-0.260 (±0.090)
White	0.513 (±0.037)	0.513 (±0.038)	0.000 (±0.053)
Black	0.778 (±0.028)	0.661 (±0.023)	-0.117 (±0.036)

Table C.4: Bias comparison between decentralized training and central training for data partition with quantity heterogeneity introduced by using all original states. Results are the average over five runs with different random seeds, and their corresponding standard deviation.

	Central	**Global**	**Difference**
Accuracy	0.657 (±0.032)	0.739 (±0.008)	+0.082 (±0.033)
DP			
Sex	0.146 (±0.067)	0.562 (±0.029)	+0.416 (±0.073)
Race	0.136 (±0.133)	0.330 (±0.047)	+0.194 (±0.141)
White	0.531 (±0.088)	0.583 (±0.038)	+0.052 (±0.096)
Black	0.518 (±0.104)	0.667 (±0.041)	+0.149 (±0.112)
EO			
Sex	0.133 (±0.043)	0.536 (±0.030)	+0.403 (±0.052)
Race	0.080 (±0.172)	0.286 (±0.086)	+0.206 (±0.192)
White	0.567 (±0.115)	0.513 (±0.037)	-0.054 (±0.121)
Black	0.639 (±0.076)	0.778 (±0.028)	+0.139 (±0.081)

Partition 4: Heterogeneity in Quantity (State-based: polarized)

Table C.5 and Table C.6 show the comparison between the global and local pipelines, and the comparison between the central and global pipelines, respectively, for the data partition with quantity heterogeneity introduced by using polarized (in state size) original states. The first comparison show a discrepancy between both pipelines for the EO and DP score of Race, and for the EO score of Black/Non-black. The second comparison show that for almost all metrics, the central pipeline is more fair compared to the global FL pipeline, although the differences seem relatively small.

Table C.5: Bias metric comparison between global bias assessment and aggregated local bias assessment for data partition with quantity heterogeneity introduced by using polarized original states. Results are the average over five runs with different random seeds, and their corresponding standard deviation.

	Global	**Local**	**Difference**
Accuracy	0.741 (±0.007)	0.741 (±0.007)	0.000 (±0.010)
DP			
Sex	0.591 (±0.022)	0.590 (±0.022)	-0.001 (±0.031)
Race	0.335 (±0.048)	0.158 (±0.040)	-0.197 (±0.062)
White	0.607 (±0.026)	0.604 (±0.026)	-0.003 (±0.037)
Black	0.713 (±0.033)	0.704 (±0.028)	-0.009 (±0.043)
EO			
Sex	0.548 (±0.020)	0.538 (±0.020)	-0.010 (±0.028)
Race	0.330 (±0.072)	0.074 (±0.026)	-0.256 (±0.077)
White	0.533 (±0.025)	0.538 (±0.024)	+0.005 (±0.035)
Black	0.808 (±0.029)	0.622 (±0.019)	-0.186 (±0.035)

Table C.6: Bias comparison between decentralized training and central training for data partition with quantity heterogeneity introduced by using polarized original states. Results are the average over five runs with different random seeds, and their corresponding standard deviation.

	Central	**Global**	**Difference**
Accuracy	0.726 (±0.051)	0.741 (±0.007)	+0.015 (±0.051)
DP			
Sex	0.657 (±0.073)	0.591 (±0.022)	-0.066 (±0.076)
Race	0.423 (±0.177)	0.335 (±0.048)	-0.068 (±0.183)
White	0.728 (±0.098)	0.607 (±0.026)	-0.121 (±0.101)
Black	0.778 (±0.097)	0.713 (±0.033)	-0.065 (±0.102)
EO			
Sex	0.616 (±0.060)	0.548 (±0.020)	-0.068 (±0.063)
Race	0.390 (±0.234)	0.330 (±0.072)	-0.060 (±0.245)
White	0.712 (±0.145)	0.533 (±0.025)	-0.179 (±0.147)
Black	0.876 (±0.066)	0.808 (±0.029)	-0.068 (±0.072)

Partition 5: Heterogeneity in Quantity (State-based: distributed)

Table C.7 and Table C.8 show the comparison between the global and local pipelines, and the comparison between the central and global pipelines, respectively, for the data partition with quantity heterogeneity introduced by using distributed (in state size) original states. The first comparison show a discrepancy between both pipelines for the EO and DP score of Race, and for the EO score of Black/Non-black. The second comparison show that the metrics with respect to Sex show much less bias in the FL pipeline compared to the central pipeline, and that the differences for the other metrics are relatively small.

Table C.7: Bias metric comparison between global bias assessment and aggregated local bias assessment for data partition with quantity heterogeneity introduced by using distributed original states. Results are the average over five runs with different random seeds, and their corresponding standard deviation.

	Global	**Local**	**Difference**
Accuracy	0.728 (±0.007)	0.728 (±0.007)	0.000 (±0.010)
DP			
Sex	0.687 (±0.077)	0.687 (±0.079)	0.000 (±0.110)
Race	0.301 (±0.026)	0.061 (±0.013)	-0.240 (±0.029)
White	0.495 (±0.024)	0.500 (±0.024)	+0.005 (±0.034)
Black	0.606 (±0.022)	0.525 (±0.023)	-0.081 (±0.032)
EO			
Sex	0.770 (±0.110)	0.729 (±0.086)	-0.041 (±0.140)
Race	0.258 (±0.052)	0.013 (±0.016)	-0.245 (±0.054)
White	0.473 (±0.025)	0.445 (±0.022)	-0.028 (±0.033)
Black	0.714 (±0.028)	0.465 (±0.028)	-0.249 (±0.040)

Table C.8: Bias comparison between decentralized training and central training for data partition with quantity heterogeneity introduced by using distributed original states. Results are the average over five runs with different random seeds, and their corresponding standard deviation.

	Central	**Global**	**Difference**
Accuracy	0.734 (±0.010)	0.728 (±0.007)	-0.006 (±0.012)
DP			
Sex	0.467 (±0.047)	0.687 (±0.077)	+0.220 (±0.090)
Race	0.331 (±0.089)	0.301 (±0.026)	-0.030 (±0.093)
White	0.557 (±0.057)	0.495 (±0.024)	-0.062 (±0.062)
Black	0.594 (±0.072)	0.606 (±0.022)	+0.012 (±0.075)
EO			
Sex	0.441 (±0.056)	0.770 (±0.110)	+0.329 (±0.123)
Race	0.252 (±0.156)	0.258 (±0.052)	+0.006 (±0.164)
White	0.574 (±0.066)	0.473 (±0.025)	-0.101 (±0.071)
Black	0.725 (±0.069)	0.714 (±0.028)	-0.011 (±0.074)

Partition 6: Heterogeneity in Label (100-0)

Table C.9 and Table C.10 show the comparison between the global and local pipelines, and the comparison between the central and global pipelines, respectively, for the data partition with full label heterogeneity (100-0). The first comparison show missing values for the EO scores of the aggregated local assessment, since every client has datasamples with the same label. The DP scores show an increase in the DP score for the White/Non-white categorization. The second comparison shows that both metrics for the Sex and Race features measure less bias in the global FL method compared to the central pipeline. The other categorizations measured more bias in the global FL model.

Table C.9: Bias metric comparison between global bias assessment and aggregated local bias assessment for data partition with full label heterogeneity (100-0). Results are the average over five runs with different random seeds, and their corresponding standard deviation.

	Global	**Local**	**Difference**
Accuracy	0.494 (±0.022)	0.494 (±0.022)	0.000 (±0.031)
DP			
Sex	0.921 (±0.110)	0.851 (±0.206)	-0.070 (±0.234)
Race	0.626 (±0.390)	0.546 (±0.424)	-0.080 (±0.576)
White	0.387 (±0.358)	0.532 (±0.307)	+0.145 (±0.472)
Black	0.500 (±0.425)	0.363 (±0.175)	-0.137 (±0.460)
EO			
Sex	0.900 (±0.116)	-	-
Race	0.600 (±0.384)	-	-
White	0.580 (±0.357)	-	-
Black	0.488 (±0.417)	-	-

Table C.10: Bias comparison between decentralized training and central training for data partition with full label heterogeneity (100-0). Results are the average over five runs with different random seeds, and their corresponding standard deviation.

	Central	**Global**	**Difference**
Accuracy	0.698 (±0.040)	0.494 (±0.022)	-0.204 (±0.046)
DP			
Sex	0.698 (±0.116)	0.921 (±0.110)	+0.223 (±0.160)
Race	0.483 (±0.232)	0.626 (±0.390)	+0.143 (±0.454)
White	0.749 (±0.116)	0.387 (±0.358)	-0.362 (±0.376)
Black	0.774 (±0.128)	0.500 (±0.425)	-0.274 (±0.444)
EO			
Sex	0.655 (±0.112)	0.900 (±0.116)	+0.245 (±0.161)
Race	0.488 (±0.225)	0.600 (±0.384)	+0.112 (±0.445)
White	0.713 (±0.139)	0.580 (±0.357)	-0.133 (±0.383)
Black	0.855 (±0.092)	0.488 (±0.417)	-0.367 (±0.427)

Partition 7: Heterogeneity in Label (75-25)

Table C.11 and Table C.12 show the comparison between the global and local pipelines, and the comparison between the central and global pipelines, respectively, for the data partition with label heterogeneity (75-25). The first comparison shows a discrepancy between both pipelines for the EO and DP score of Race, and for the EO score of Black/Non-black. The second comparison shows that the EO and DP metric for Sex measure less biases in the global FL method compared to the central pipeline. The other categorizations measured more bias in the global FL model.

Table C.11: Bias metric comparison between global bias assessment and aggregated local bias assessment for data partition with label heterogeneity (75-25). Results are the average over five runs with different random seeds, and their corresponding standard deviation.

	Global	**Local**	**Difference**
Accuracy	0.737 (±0.003)	0.737 (±0.003)	0.000 (±0.004)
DP			
Sex	0.741 (±0.062)	0.758 (±0.064)	+0.017 (±0.089)
Race	0.405 (±0.024)	0.116 (±0.019)	-0.289 (±0.031)
White	0.651 (±0.014)	0.661 (±0.014)	+0.010 (±0.020)
Black	0.731 (±0.018)	0.753 (±0.018)	+0.022 (±0.025)
EO			
Sex	0.777 (±0.109)	0.763 (±0.097)	-0.014 (±0.146)
Race	0.392 (±0.014)	0.017 (±0.008)	-0.375 (±0.016)
White	0.569 (±0.011)	0.570 (±0.013)	+0.001 (±0.017)
Black	0.803 (±0.008)	0.729 (±0.023)	-0.074 (±0.024)

Table C.12: Bias comparison between decentralized training and central training for data partition with label heterogeneity (75-25). Results are the average over five runs with different random seeds, and their corresponding standard deviation.

	Central	**Global**	**Difference**
Accuracy	0.698 (±0.056)	0.737 (±0.003)	-0.039 (±0.056)
DP			
Sex	0.806 (±0.119)	0.741 (±0.062)	-0.065 (±0.134)
Race	0.687 (±0.195)	0.405 (±0.024)	-0.282 (±0.196)
White	0.853 (±0.131)	0.651 (±0.014)	-0.202 (±0.132)
Black	0.893 (±0.118)	0.731 (±0.018)	-0.162 (±0.119)
EO			
Sex	0.655 (±0.115)	0.777 (±0.109)	+0.122 (±0.158)
Race	0.488 (±0.236)	0.392 (±0.014)	-0.096 (±0.236)
White	0.713 (±0.169)	0.569 (±0.011)	-0.144 (±0.169)
Black	0.855 (±0.373)	0.803 (±0.008)	-0.052 (±0.373)

Partition 8: Heterogeneity in Feature (Sex, 100-0)

Table C.13 and Table C.14 show the comparison between the global and local pipelines, and the comparison between the central and global pipelines, respectively, for the data partition with feature heterogeneity for Sex (100-0). The first comparison show missing values for the Sex attribute, due to a single sex group in every client. Furthermore, a discrepancy between both pipelines for the EO and DP score of Race, and for the EO score of Black/Non-black is visible. The second comparison shows that the EO and DP metric for Sex measure less bias in the global FL method compared to the central pipeline. The other categorizations measured more bias in the global FL model.

Table C.13: Bias metric comparison between global bias assessment and aggregated local bias assessment for data partition with feature heterogeneity in Sex (100-0). Results are the average over five runs with different random seeds, and their corresponding standard deviation.

	Global	**Local**	**Difference**
Accuracy	0.727 (±0.004)	0.727 (±0.004)	0.000 (±0.006)
DP			
Sex	0.866 (±0.112)	-	-
Race	0.326 (±0.008)	0.048 (±0.014)	-0.278 (±0.016)
White	0.584 (±0.008)	0.584 (±0.008)	0.000 (±0.011)
Black	0.690 (±0.014)	0.695 (±0.009)	+0.005 (±0.017)
EO			
Sex	0.822 (±0.085)	-	-
Race	0.308 (±0.015)	0.001 (±0.002)	-0.307 (±0.015)
White	0.491 (±0.015)	0.494 (±0.013)	+0.003 (±0.020)
Black	0.777 (±0.009)	0.726 (±0.014)	-0.051 (±0.017)

Table C.14: Bias comparison between decentralized training and central training for data partition with feature heterogeneity in *Sex*(100-0). Results are the average over five runs with different random seeds, and their corresponding standard deviation.

	Central	**Global**	**Difference**
Accuracy	0.698 (±0.062)	0.727 (±0.004)	+0.029 (±0.062)
DP			
Sex	0.650 (±0.079)	0.866 (±0.112)	+0.216 (±0.137)
Race	0.413 (±0.170)	0.326 (±0.008)	-0.087 (±0.170)
White	0.714 (±0.086)	0.584 (±0.008)	-0.130 (±0.086)
Black	0.750 (±0.097)	0.690 (±0.014)	-0.060 (±0.098)
EO			
Sex	0.655 (±0.069)	0.822 (±0.085)	+0.167 (±0.109)
Race	0.488 (±0.169)	0.308 (±0.015)	-0.180 (±0.170)
White	0.713 (±0.110)	0.491 (±0.015)	-0.222 (±0.111)
Black	0.855 (±0.065)	0.777 (±0.009)	-0.078 (±0.066)

Partition 9: Heterogeneity in Feature (Sex, 75-25)

Table C.15 and Table C.16 show the comparison between the global and local pipelines, and the comparison between the central and global pipelines, respectively, for the data partition with feature heterogeneity for Sex (75-25). The first comparison shows a discrepancy between both pipelines for the EO and DP score of Race, and for the EO score of Black/Non-black is visible. The second comparison shows that the EO and DP metric for Sex measure less bias in the global FL method compared to the central pipeline. The other categorizations measured more bias in the global FL model.

Table C.15: Bias metric comparison between global bias assessment and aggregated local bias assessment for data partition with feature heterogeneity in Sex (75-25). Results are the average over five runs with different random seeds, and their corresponding standard deviation.

	Global	**Local**	**Difference**
Accuracy	0.734 (±0.004)	0.734 (±0.004)	0.000 (±0.006)
DP			
Sex	0.724 (±0.057)	0.724 (±0.058)	0.000 (±0.025)
Race	0.344 (±0.017)	0.047 (±0.018)	-0.297 (±0.157)
White	0.588 (±0.020)	0.588 (±0.020)	0.000 (±0.028)
Black	0.676 (±0.021)	0.679 (±0.021)	+0.003 (±0.030)
EO			
Sex	0.795 (±0.096)	0.773 (±0.082)	-0.022 (±0.126)
Race	0.308 (±0.038)	0.007 (±0.005)	-0.301 (±0.038)
White	0.516 (±0.029)	0.517 (±0.028)	+0.001 (±0.040)
Black	0.775 (±0.025)	0.700 (±0.030)	-0.075 (±0.039)

Table C.16: Bias comparison between decentralized training and central training for data partition with feature heterogeneity in Sex(75-25). Results are the average over five runs with different random seeds, and their corresponding standard deviation.

	Central	**Global**	**Difference**
Accuracy	0.687 (±0.053)	0.734 (±0.004)	+0.047 (±0.053)
DP			
Sex	0.690 (±0.080)	0.724 (±0.057)	+0.034 (±0.098)
Race	0.506 (±0.157)	0.344 (±0.017)	-0.162 (±0.158)
White	0.768 (±0.105)	0.588 (±0.020)	-0.180 (±0.107)
Black	0.815 (±0.115)	0.676 (±0.021)	-0.139 (±0.117)
EO			
Sex	0.644 (±0.069)	0.795 (±0.096)	+0.151 (±0.118)
Race	0.474 (±0.181)	0.308 (±0.038)	-0.166 (±0.185)
White	0.757 (±0.133)	0.516 (±0.029)	-0.241 (±0.136)
Black	0.890 (±0.085)	0.775 (±0.025)	-0.115 (±0.089)

Partition 10: Heterogeneity in Feature (Race)

Table C.17 and Table C.18 show the comparison between the global and local pipelines, and the comparison between the central and global pipelines, respectively, for the data partition with feature heterogeneity in Race. The first comparison shows missing values for the Race attribute, due to the individual race-label per client. Furthermore, a discrepancy between both pipelines for the EO and DP score of the Black/Non-black categorization is visible. The second comparison shows that the EO and DP metric for Sex and Race measure more bias in the global FL method compared to the central pipeline. The White/Non-white categorizations measured less bias in the global FL model.

Table C.17: Bias metric comparison between global bias assessment and aggregated local bias assessment for data partition with feature heterogeneity in RACE. Results are the average over five runs with different random seeds, and their corresponding standard deviation.

	Global	**Local**	**Difference**
Accuracy	0.658 (±0.009)	0.658 (±0.009)	0.000 (±0.013)
DP			
Sex	0.603 (±0.262)	0.765 (±0.146)	+0.162 (±0.300)
Race	0.017 (±0.033)	-	-
White	0.678 (±0.245)	0.889 (±0.000)	+0.211 (±0.245)
Black	0.348 (±0.426)	0.111 (±0.000)	-0.237 (±0.426)
EO			
Sex	0.342 (±0.238)	0.748 (±0.176)	+0.406 (±0.296)
Race	0.000 (±0.000)	-	-
White	0.781 (±0.270)	0.889 (±0.000)	+0.108 (±0.270)
Black	0.162 (±0.324)	0.111 (±0.000)	-0.051 (±0.324)

Table C.18: Bias comparison between decentralized training and central training for data partition with feature heterogeneity in RACE. Results are the average over five runs with different random seeds, and their corresponding standard deviation.

	Central	**Global**	**Difference**
Accuracy	0.666 (±0.095)	0.658 (±0.009)	-0.008 (±0.095)
DP			
Sex	0.904 (±0.063)	0.603 (±0.262)	-0.301 (±0.269)
Race	0.392 (±0.150)	0.017 (±0.033)	-0.375 (±0.154)
White	0.584 (±0.151)	0.678 (±0.245)	+0.094 (±0.288)
Black	0.273 (±0.354)	0.348 (±0.426)	+0.075 (±0.554)
EO			
Sex	0.852 (±0.063)	0.342 (±0.238)	-0.510 (±0.246)
Race	0.283 (±0.155)	0.000 (±0.000)	-0.283 (±0.155)
White	0.570 (±0.152)	0.781 (±0.270)	+0.211 (±0.310)
Black	0.396 (±0.287)	0.162 (±0.324)	-0.234 (±0.433)

Partition 11: Heterogeneity in Feature (Marital Status)

Table C.19 and Table C.20 show the comparison between the global and local pipelines, and the comparison between the central and global pipelines, respectively, for the data partition with feature heterogeneity in MARITAL STATUS. The first comparison shows discrepancies between both pipelines across all metrics. The second comparison shows that the DP score for SEX and the EO score for WHITE/NON-WHITE show less bias in the global FL model, but that the other metrics measured slightly more bias. However, these differences are quite small.

Table C.19: Bias metric comparison between global bias assessment and aggregated local bias assessment for data partition with feature heterogeneity in Marital Status. Results are the average over five runs with different random seeds, and their corresponding standard deviation.

	Global	**Local**	**Difference**
Accuracy	0.543 (±0.047)	0.543 (±0.047)	0.000 (±0.066)
DP			
Sex	0.669 (±0.145)	0.688 (±0.153)	+0.019 (±0.211)
Race	0.197 (±0.106)	0.058 (±0.023)	-0.139 (±0.108)
White	0.645 (±0.139)	0.579 (±0.074)	-0.066 (±0.157)
Black	0.589 (±0.320)	0.496 (±0.052)	-0.093 (±0.324)
EO			
Sex	0.547 (±0.197)	0.504 (±0.199)	-0.043 (±0.280)
Race	0.082 (±0.106)	0.005 (±0.006)	-0.077 (±0.106)
White	0.675 (±0.099)	0.614 (±0.119)	-0.061 (±0.155)
Black	0.625 (±0.231)	0.423 (±0.076)	-0.202 (±0.243)

Table C.20: Bias comparison between decentralized training and central training for data partition with feature heterogeneity in Marital Status. Results are the average over five runs with different random seeds, and their corresponding standard deviation.

	Central	**Global**	**Difference**
Accuracy	0.647 (±0.062)	0.543 (±0.047)	-0.104 (±0.078)
DP			
Sex	0.614 (±0.132)	0.669 (±0.145)	+0.055 (±0.196)
Race	0.274 (±0.235)	0.197 (±0.106)	-0.077 (±0.258)
White	0.689 (±0.089)	0.645 (±0.139)	-0.044 (±0.165)
Black	0.869 (±0.115)	0.589 (±0.320)	-0.280 (±0.340)
EO			
Sex	0.566 (±0.120)	0.547 (±0.197)	-0.019 (±0.231)
Race	0.175 (±0.151)	0.082 (±0.106)	-0.093 (±0.184)
White	0.649 (±0.100)	0.675 (±0.099)	+0.026 (±0.141)
Black	0.700 (±0.333)	0.625 (±0.231)	-0.075 (±0.405)

Partition 12: Heterogeneity in Label (100-0) + Quantity

Table C.21 and Table C.22 show the comparison between the global and local pipelines, and the comparison between the central and global pipelines, respectively, for the data partition with heterogeneity in label (100-0) and quantity. The first comparison shows higher discrepancies for the DP metrics of Race and White/Non-white. The second comparison shows that the DP and EO score for Sex and White/Non-white show less bias in the global FL model, but that the other metrics measured slightly more bias.

Table C.21: Bias metric comparison between global bias assessment and aggregated local bias assessment for data partition with heterogeneity in label (100-0) and quantity. Results are the average over five runs with different random seeds, and their corresponding standard deviation.

	Global	**Local**	**Difference**
Accuracy	0.708 (±0.182)	0.707 (±0.181)	-0.001 (±0.257)
DP			
Sex	0.751 (±0.180)	0.710 (±0.197)	-0.041 (±0.267)
Race	0.058 (±0.117)	0.396 (±0.356)	+0.338 (±0.375)
White	0.656 (±0.167)	0.390 (±0.240)	-0.266 (±0.292)
Black	0.427 (±0.269)	0.481 (±0.297)	+0.054 (±0.401)
EO			
Sex	0.594 (±0.313)	-	-
Race	0.038 (±0.077)	-	-
White	0.756 (±0.204)	-	-
Black	0.578 (±0.320)	-	-

Table C.22: Bias comparison between decentralized training and central training for data partition with heterogeneity in label (100-0) and quantity. Results are the average over five runs with different random seeds, and their corresponding standard deviation.

	Central	**Global**	**Difference**
Accuracy	0.826 (±0.007)	0.708 (±0.182)	-0.118 (±0.182)
DP			
Sex	0.262 (±0.071)	0.751 (±0.180)	+0.489 (±0.193)
Race	0.166 (±0.035)	0.058 (±0.117)	-0.108 (±0.122)
White	0.517 (±0.113)	0.656 (±0.167)	+0.139 (±0.202)
Black	0.546 (±0.090)	0.427 (±0.269)	-0.119 (±0.284)
EO			
Sex	0.256 (±0.066)	0.594 (±0.313)	+0.338 (±0.320)
Race	0.069 (±0.091)	0.038 (±0.077)	-0.031 (±0.119)
White	0.521 (±0.130)	0.756 (±0.204)	+0.235 (±0.242)
Black	0.665 (±0.092)	0.578 (±0.320)	-0.087 (±0.333)

Partition 13: Heterogeneity in Label (75-25) + Quantity

Table C.23 and Table C.24 show the comparison between the global and local pipelines, and the comparison between the central and global pipelines, respectively, for the data partition with heterogeneity in label (75-25) and quantity. The first comparison shows a higher discrepancy for the DP metric of Race. The second comparison shows that the DP and EO score for Sex metric and the EO score for Black/Non-black measure less bias in the global FL model.

Table C.23: Bias metric comparison between global bias assessment and aggregated local bias assessment for data partition with heterogeneity in label (75-25) and quantity. Results are the average over five runs with different random seeds, and their corresponding standard deviation.

	Global	**Local**	**Difference**
Accuracy	0.707 (±0.011)	0.707 (±0.011)	0.000 (±0.016)
DP			
Sex	0.909 (±0.019)	0.895 (±0.026)	-0.014 (±0.032)
Race	0.446 (±0.131)	0.124 (±0.055)	-0.322 (±0.142)
White	0.687 (±0.095)	0.695 (±0.098)	+0.008 (±0.136)
Black	0.808 (±0.027)	0.724 (±0.036)	-0.084 (±0.045)
EO			
Sex	0.788 (±0.119)	-	-
Race	0.386 (±0.174)	-	-
White	0.619 (±0.159)	-	-
Black	0.871 (±0.031)	-	-

Table C.24: Bias comparison between decentralized training and central training for data partition with heterogeneity in label (75-25) and quantity. Results are the average over five runs with different random seeds, and their corresponding standard deviation.

	Central	**Global**	**Difference**
Accuracy	0.691 (±0.051)	0.707 (±0.011)	+0.016 (±0.052)
DP			
Sex	0.704 (±0.139)	0.909 (±0.019)	+0.205 (±0.140)
Race	0.481 (±0.238)	0.446 (±0.131)	-0.035 (±0.272)
White	0.741 (±0.152)	0.687 (±0.095)	-0.054 (±0.179)
Black	0.817 (±0.129)	0.808 (±0.027)	-0.009 (±0.132)
EO			
Sex	0.639 (±0.138)	0.788 (±0.119)	+0.149 (±0.182)
Race	0.364 (±0.274)	0.386 (±0.174)	+0.022 (±0.325)
White	0.676 (±0.195)	0.619 (±0.159)	-0.057 (±0.252)
Black	0.689 (±0.347)	0.871 (±0.031)	+0.182 (±0.348)

Partition 14: Heterogeneity in Feature (Sex, 100-0) + Quantity

Table C.25 and Table C.26 show the comparison between the global and local pipelines, and the comparison between the central and global pipelines, respectively, for the data partition with heterogeneity in feature Sex (100-0) and quantity. The first comparison show missing values for the metric with respect to Sex, since every client has only data samples with the sample sex-label. Furthermore, this comparison shows higher discrepancy for the DP and EO metrics for Race and the EO metric for White/Non-white. The second comparison shows less biases in the global FL model with respect Sex as measure by the DP metric. There is more bias with respect to Race and Black/Non-black as measured by both the DP and EO metrics.

Table C.25: Bias metric comparison between global bias assessment and aggregated local bias assessment for data partition with heterogeneity in feature Sex (100-0) and quantity. Results are the average over five runs with different random seeds, and their corresponding standard deviation.

	Global	**Local**	**Difference**
Accuracy	0.718 (±0.007)	0.718 (±0.007)	0.000 (±0.010)
DP			
Sex	0.847 (±0.077)	-	-
Race	0.305 (±0.025)	0.101 (±0.025)	-0.204 (±0.035)
White	0.555 (±0.017)	0.552 (±0.015)	-0.003 (±0.023)
Black	0.682 (±0.041)	0.648 (±0.023)	-0.034 (±0.047)
EO			
Sex	0.712 (±0.149)	-	-
Race	0.275 (±0.040)	0.010 (±0.007)	-0.265 (±0.041)
White	0.471 (±0.030)	0.483 (±0.031)	+0.012 (±0.043)
Black	0.774 (±0.040)	0.625 (±0.030)	-0.149 (±0.050)

Table C.26: Bias comparison between decentralized training and central training for data partition with heterogeneity in feature Sex (100-0) and quantity. Results are the average over five runs with different random seeds, and their corresponding standard deviation.

	Central	**Global**	**Difference**
Accuracy	0.667 (±0.078)	0.718 (±0.007)	+0.051 (±0.078)
DP			
Sex	0.735 (±0.103)	0.847 (±0.077)	+0.112 (±0.129)
Race	0.581 (±0.242)	0.305 (±0.025)	-0.276 (±0.243)
White	0.777 (±0.122)	0.555 (±0.017)	-0.222 (±0.123)
Black	0.849 (±0.118)	0.682 (±0.041)	-0.167 (±0.125)
EO			
Sex	0.706 (±0.098)	0.712 (±0.149)	+0.006 (±0.178)
Race	0.571 (±0.219)	0.275 (±0.040)	-0.296 (±0.223)
White	0.753 (±0.141)	0.471 (±0.030)	-0.282 (±0.144)
Black	0.718 (±0.361)	0.774 (±0.040)	+0.056 (±0.363)

Partition 15: Heterogeneity in Feature (Sex, 75-25) + Quantity

Table C.27 and Table C.28 show the comparison between the global and local pipelines, and the comparison between the central and global pipelines, respectively, for the data partition with heterogeneity in feature Sex (75-25) and quantity. The first comparison shows higher discrepancies for the DP and EO score with respect to Race and the EO score with respect to Black/Non-black. The second comparison shows less bias in the global FL model with respect Sex as measured by both the DP metric and EO metric.

Table C.27: Bias metric comparison between global bias assessment and aggregated local bias assessment for data partition with heterogeneity in feature Sex(75-25) and quantity. Results are the average over five runs with different random seeds, and their corresponding standard deviation.

	Global	**Local**	**Difference**
Accuracy	0.717 (±0.009)	0.717 (±0.009)	0.000 (±0.013)
DP			
Sex	0.891 (±0.083)	0.872 (±0.066)	-0.019 (±0.106)
Race	0.271 (±0.088)	0.047 (±0.022)	-0.224 (±0.091)
White	0.590 (±0.081)	0.591 (±0.080)	+0.001 (±0.114)
Black	0.710 (±0.055)	0.661v (±0.040)	-0.049 (±0.068)
EO			
Sex	0.807 (±0.079)	0.762 (±0.051)	-0.045 (±0.094)
Race	0.230 (±0.146)	0.014 (±0.010)	-0.216 (±0.146)
White	0.518 (±0.109)	0.514 (±0.098)	+0.004 (±0.147)
Black	0.796 (±0.048)	0.575 (±0.049)	-0.221 (±0.069)

Table C.28: Bias comparison between decentralized training and central training for data partition with heterogeneity in feature Sex (75-25) and quantity. Results are the average over five runs with different random seeds, and their corresponding standard deviation.

	Central	**Global**	**Difference**
Accuracy	0.714 (±0.013)	0.717 (±0.009)	+0.003 (±0.016)
DP			
Sex	0.509 (±0.044)	0.891 (±0.083)	+0.382 (±0.094)
Race	0.251 (±0.097)	0.271 (±0.088)	+0.020 (±0.131)
White	0.667 (±0.075)	0.590 (±0.081)	-0.077 (±0.110)
Black	0.644 (±0.067)	0.710 (±0.055)	+0.066 (±0.087)
EO			
Sex	0.499 (±0.051)	0.807 (±0.079)	+0.308 (±0.094)
Race	0.170 (±0.153)	0.230 (±0.146)	+0.060 (±0.211)
White	0.616 (±0.117)	0.518 (±0.109)	-0.098 (±0.160)
Black	0.720 (±0.126)	0.796 (±0.048)	+0.076 (±0.135)

Partition 16: Heterogeneity in Feature (Race) + Quantity

Table C.29 and Table C.30 show the comparison between the global and local pipelines, and the comparison between the central and global pipelines, respectively, for the data partition with heterogeneity in feature Race and quantity. The first comparison shows missing values for the local measured biases with respect to Race, since every client has only data samples from the same race group. Furthermore, this comparison shows higher discrepancy for the DP and EO metrics for White/Non-white, where the local pipeline measured less bias. The second comparison shows less bias in the global FL model with respect Sex as measured by both demographic parity and equalized odds. There is more bias in the global FL pipeline with respect to Race and White/Non-white as measured by both the DP and EO metrics compared to the central pipeline.

Table C.29: Bias metric comparison between global bias assessment and aggregated local bias assessment for data partition with heterogeneity in feature RACE and quantity. Results are the average over five runs with different random seeds, and their corresponding standard deviation.

	Global	**Local**	**Difference**
Accuracy	0.636 (±0.024)	0.636 (±0.024)	0.000 (±0.034)
DP			
Sex	0.835 (±0.033)	0.814 (±0.032)	-0.021 (±0.046)
Race	0.000 (±0.000)	-	-
White	0.142 (±0.009)	0.331 (±0.000)	+0.189 (±0.009)
Black	0.739 (±0.044)	0.083 (±0.000)	-0.656 (±0.044)
EO			
Sex	0.843 (±0.035)	0.785 (±0.035)	-0.058 (±0.049)
Race	0.000 (±0.000)	-	-
White	0.119 (±0.010)	0.331 (±0.000)	+0.212 (±0.010)
Black	0.628 (±0.051)	0.083 (±0.000)	-0.545 (±0.051)

Table C.30: Bias comparison between decentralized training and central training for data partition with heterogeneity in feature RACE and quantity. Results are the average over five runs with different random seeds, and their corresponding standard deviation.

	Central	**Global**	**Difference**
Accuracy	0.716 (±0.055)	0.636 (±0.024)	-0.080 (±0.060)
DP			
Sex	0.607 (±0.092)	0.835 (±0.033)	+0.228 (±0.098)
Race	0.344 (±0.149)	0.000 (±0.000)	-0.344 (±0.149)
White	0.662 (±0.110)	0.142 (±0.009)	-0.520 (±0.110)
Black	0.713 (±0.110)	0.739 (±0.044)	+0.026 (±0.118)
EO			
Sex	0.577 (±0.079)	0.843 (±0.035)	+0.266 (±0.086)
Race	0.288 (±0.199)	0.000 (±0.000)	-0.288 (±0.199)
White	0.628 (±0.128)	0.119 (±0.010)	-0.509 (±0.128)
Black	0.824 (±0.077)	0.628 (±0.051)	-0.196 (±0.092)

Partition 17: Heterogeneity in Feature (Marital Status) + Quantity

Table C.31 and Table C.32 show the comparison between the global and local pipelines, and the comparison between the central and global pipelines, respectively, for the data partition with heterogeneity in feature RACE and quantity. The first comparison shows much lower bias scores for the local pipeline for SEX, RACE and BLACK/NON-BLACK. The second comparison shows less bias in the global FL model with respect SEX compared to the central pipeline.

Table C.31: Bias metric comparison between global bias assessment and aggregated local bias assessment for data partition with heterogeneity in feature Marital Status and quantity. Results are average over five runs with different random seeds, and their corresponding stand devitation. Results are the average over five runs with different random seeds, and their corresponding standard deviation.

	Global	**Local**	**Difference**
Accuracy	0.708 (±0.009)	0.708 (±0.009)	0.000 (±0.013)
DP			
Sex	0.757 (±0.030)	0.548 (±0.020)	-0.209 (±0.036)
Race	0.438 (±0.074)	0.264 (±0.053)	-0.174 (±0.091)
White	0.665 (±0.025)	0.653 (±0.025)	-0.012 (±0.035)
Black	0.674 (±0.013)	0.635 (±0.044)	-0.039 (±0.046)
EO			
Sex	0.839 (±0.029)	0.486 (±0.025)	-0.353 (±0.038)
Race	0.422 (±0.090)	0.107 (±0.037)	-0.315 (±0.097)
White	0.622 (±0.033)	0.595 (±0.030)	-0.027 (±0.045)
Black	0.726 (±0.017)	0.425 (±0.081)	-0.301 (±0.083)

Table C.32: Bias comparison between decentralized training and central training for data partition with heterogeneity in feature Marital Status and quantity. Results are the average over five runs with different random seeds, and their corresponding standard deviation.

	Central	**Global**	**Difference**
Accuracy	0.742 (±0.011)	0.708 (±0.009)	-0.034 (±0.014)
DP			
Sex	0.602 (±0.041)	0.757 (±0.030)	+0.155 (±0.051)
Race	0.330 (±0.112)	0.438 (±0.074)	+0.108 (±0.134)
White	0.686 (±0.062)	0.665 (±0.025)	-0.021 (±0.067)
Black	0.723 (±0.072)	0.674 (±0.013)	-0.049 (±0.073)
EO			
Sex	0.567 (±0.025)	0.839 (±0.029)	+0.272 (±0.038)
Race	0.329 (±0.147)	0.422 (±0.090)	+0.093 (±0.172)
White	0.667 (±0.091)	0.622 (±0.033)	-0.045 (±0.097)
Black	0.835 (±0.054)	0.726 (±0.033)	-0.109 (±0.057)