Vrije Universiteit Amsterdam

Universiteit van Amsterdam

Master Thesis

Investigation on the Effect of Data Bias in Differentially Private Federated Learning

Author: Yu Wang (2646423 / 12962988)

1st supervisor: Adam Belloum
daily supervisor: Saba Amiri
2nd reader: Zhiming Zhao

A thesis submitted in fulfillment of the requirements for the joint UvA-VU Master of Science degree in Computer Science

August 31, 2021

"I am the master of my fate, I am the captain of my soul"
from Invictus*, by William Ernest Henley*

Abstract

Privacy-utility tradeoff and fairness on under-represented groups have become central issues in a differentially private federated learning framework. This thesis aims at measuring the utility and fairness of a distributed learning system, by investigating the effect of data bias and privacy preserving mechanism within a federated learning setting.

This thesis considers data bias from two aspects:(1) data imbalance within a dataset and (2) non-IID data distribution among worker nodes. Specifically, the effect of data bias is measured in three parts: (1) target class imbalance, (2) label imbalance, and (3) imbalanced number of samples among worker nodes.

This thesis designed and implemented a comprehensive experiment scheme for measuring the effect of data bias in differentially private federated learning. This thesis simulated 4 types of representative data distributions scenarios based on real-world machine learning problems: (1) fully IID, (2) partial 2-class non-IID, (3) fully 2-class non-IID, and (4) normal distribution. This thesis chose 9 privacy budgets (ϵ value) from 0.2 to 100 to simulate different privacy level required by the worker nodes itself or the legislation.

This thesis conducted 60 experiments with 6 data distribution scenarios and 10 differential privacy settings. Vertical and horizontal comparisons among experiment results are performed to validate the following hypotheses: (1) a higher level of data bias leads to a better overall performance and worse fairness of under-represented groups, and (2) a higher privacy level leads to a worse overall performance and worse fairness of under-represented groups.

This thesis found that there is a large performance difference between the target class with most samples and the target class with only a few samples in a classification task on a highly imbalanced dataset. Also, the model

performance of minority groups is significantly influenced by the changes of differentially private federated learning setting, compared with the overall dataset and majority groups.

Acknowledgements

The author of this thesis would like to thank Zhijun Liu (Susan) for her great support and love. The author wants to express her gratitude to Adam Belloum and Saba Amiri for their guidance and help in this thesis project. Also, the author would like to thank Carlijn Nijhuis and Simon Tokloth for the great discussions about the project.

Contents

List of Figures

LIST OF FIGURES

LIST OF FIGURES

List of Tables

LIST OF TABLES

1

Introduction

1.1 Context

Driven by the awareness of personal data privacy protection and stricter legislation of commercial data exchange such as General Data Protection Regulation (GDPR) in EU (1) and Health Insurance Portability and Accountability Act (HIPAA) in US (2), privacy-utility tradeoff has become a primary concern of distributed Machine Learning in the past decade. Collaboration on Machine Learning model training among various entities have shown its great demand of a privacy-preserving machine learning framework. The type of involved entities are different under various scenarios, such as the huge amount of end users for a smart device or application, first-hand collected data organized by different research institutions, internal innovative projects among different departments within one company, and different branches of an international corporate.

With this context, Federated Learning (FL) was first introduced in 2016 as a decentralized approach to leave the training data distributed on the local nodes and learn a shared model based on the aggregation of locally-computed updates on the server node (3). Later in 2019, a workshop focused on Federated Learning and Analytics was held by Google, in which researchers broadened the definition of federated learning and systematically categorized open problems in the field (4).

In the categorization of federated learning, the most significant difference between cross-device federated learning and cross-silo federated learning is the type and amount of involved entities. There is a huge amount of unreliable devices in cross-device federated learning, but the cross-silo federated learning only involves a small amount of

reliable organizations. On the basis of these two terms, the definition of federated learning was broadened to become more applicable in different types of real-world machine learning problem (4).

Aside from the higher willingness to collaborate in machine learning tasks, the quantity and quality of data have also been largely improved. In the last few decades, the development of sensor networks has extended its applicability in various fields (5), including environmental monitoring based on Internet of Things (IoT) sensors (6), human activity recognition based on image sensors (7), and disease prediction based on medical signal sensors (8). Moreover, the global trend of digital transformation in all business industries have also boosted the amount of data being generated every second. For example, YouTue uses billions of user browsing history and user persona to improve their recommendation system (9), and Alibaba uses millions of transactions to detect fraud (10).

Aforementioned rich data have prompted a surge of interest in utilizing machine learning techniques to solve real-world problems. These tasks are sometimes interdisciplinary applications, ranging from personal credit score prediction using bank transactions and webshop browsing history (11) to length of stay in Intensive Care Units (ICU) prediction using hospital patient visit records and inventory list (12).

Given that real-world machine learning tasks usually involve multiple parties in a complex business setting as mentioned above, they need information from multiple data sources owned by different entities. These data sources have the following characteristics: (1) variety in sensitivity of the data itself, (2) difference in legislation among geographical regions, and (3) bottleneck in the communication efficiency of different data centers. Besides the general differences among various data sources, the performance of some deep neural networks is highly depended on the availability of large-scale and highly-representative datasets (13). With all these constraints, making the use of modest privacy-preserving mechanisms has become a central issue when training Machine Learning models in a distributed setting with multiple entities.

In order to train deep neural networks under a proper privacy budget to ensure the balance between model quality and data privacy, differential privacy integrated with federated learning was proposed as an appropriate solution to collaboratively train a model but keeping the data distributed in local entities (13). Within differential privacy mechanisms, the prevalent Gaussian mechanism protects data privacy by adding

elaborate Gaussian (normal) distribution as the noise to a mapping function. The noise is calibrated to the sensitivity of that mapping function from database to real values (14).

Based on the strong need of a privacy-preserving distributed machine learning framework, differentially private federated learning setting has received increasing attention because: (1) each local entity (the owner of a data source that participates in the collaborative training process) is independent and only needs to communicate model parameters (not data) with the central server, (2) the central server coordinates the aggregation and broadcast of model parameters, and (3) the noise adding mechanism within differential privacy offers a great balance between data privacy level, training efficiency, and model performance (14).

1.2 Motivation

When using the differentially private federated learning framework to solve real-world problems, data bias is widely regarded as the major limitation of the utility and fairness (4). Being different from famous Machine Learning dataset in use like MNIST (15), ImageNet (16), and movie review (17), the complicated business background of real-world Machine Learning problems makes them obliged to deal with highly imbalanced dataset with large data bias. In this thesis, we take the classification task within a privacy-preserving distributed machine learning setting as an example. With this context, we consider data bias from two aspects: (1) data imbalance within the dataset and (2) data distribution scheme which is not independent and identical among entities (non-IID data distribution). Specifically, there are three types of data bias in classification tasks under differentially private federated learning framework: (1) imbalanced number of samples within each target class, (2) imbalanced number of samples within each label, and (3) imbalanced number of samples among local entities (worker nodes).

Moreover, the data bias issue has even larger impact in situations where under-represented classes or labels within the dataset are more useful for training a Machine Learning model. For example, in the beginning of the coronavirus pandemic, the infected patients' chest X-ray images are highly under-represented classes, known as "long tails" in the whole chest X-ray image dataset. But those are the ones with greatest importance in the image classification model for pre-diagnoses of Covid-19 (18). This

scenario has the following characteristics: (1) the number of total (all-type) patients are different in each hospital, (2) the number of patients within each disease class are largely different in the whole dataset, (3) the total number of patients with Covid-19 is really small in the dataset, and (4) the number of patients with Covid-19 in each hospital varies a lot. With constraints (2) and (3), training a medical image classification model with high accuracy on Covid-19 is already really hard. Moreover, the high sensitivity level of medical data does not allow hospitals to share the raw data of their patients. Together with constraints (1) and (4), the situation is much harder for the collaboratively trained medical-use classification model. Thus, a mechanism to measure the effect of data bias is of great importance in federated learning with differential privacy.

1.3 Problem Statement

Most of the recent studies in federated learning and differential privacy are assuming that the dataset is in independently identically distribution (IID) among all worker nodes, but the real-world machine learning tasks are not always that ideal. Hence, learning the effect of data bias in differentially private federated learning framework is of great significance and gives it higher practicability and reliability in real-world machine learning tasks.

This thesis aims at measuring the utility and fairness of a distributed machine learning system, by investigating the impact of data bias and privacy preserving mechanism within a federated learning setting. We measure the effect of data bias (data imbalance and non-IID data distribution) specifically from the following three aspects in this thesis: (1) target class imbalance, (2) label imbalance, and (3) imbalanced number of samples on each worker node. The utility can be shown by measuring the model performance of the overall dataset. And the fairness of a model can be represented by the model performance of different subgroups within the whole dataset. Moreover, this thesis conducts experiments to validate the following two hypotheses: (1) a higher level of data bias leads to a better overall performance and worse fairness of under-represented groups, and (2) a higher privacy level leads to a worse overall performance and worse fairness of under-represented groups.

As the scope of this thesis, it puts attention on cross-silo federated learning setting, in which a small number of reliable organizations are involved as entities (local worker nodes). This thesis also uses the horizontal federated learning scheme, in which samples on each worker node share the same feature space but with different sample ID. Moreover, this thesis focuses on supervised learning, specifically classification tasks in machine learning problems, under differentially private federated learning framework.

1.4 Contribution

Our contributions are as follows:

- We implemented a differentially private federated learning framework with data distribution module and testing metrics module for highly imbalanced dataset, enabling the simulation of various data distribution scenarios and performance measurement of different subgroups of the data.
- We designed a comprehensive experiment scheme for measuring effect of data bias in differentially private federated learning, providing a quantitative method to systematically investigate the effect of data imbalance and non-IID data distribution.
- We conducted series of experiments on a highly imbalanced dataset and performed in-depth analysis among all experiments results, revealing the impact of various data distribution scenarios and different privacy budgets on the utility and fairness of a differentially private federated learning framework.

1.5 Thesis Structure

This thesis is structured as below:

- Chapter 2 provides previous related works about Federated Learning, Differential Privacy, data imbalance, and non-IID data distribution.
- Chapter 3 explains the data bias in a distributed machine learning task, specifically the data imbalance and non-IID data distribution.
- Chapter 4 describes the experiment design for measuring the effect of data bias in differentially private federated learning setting.

- Chapter 5 analyzes the experiment results and validates hypotheses about the impact of the higher level of data bias and higher privacy level.
- Chapter 6 shows the conclusion of this thesis.
- Chapter 7 looks forward to the directions of our future work.

2

Background

Considering that the purpose of this thesis is to measure the effect of data bias in differentially private federated learning, it is important to explain main concepts in the field of Differential Privacy (DP), Federated Learning (FL), and data bias with the context of this thesis. This chapter introduces definitions, categorizations, and current research findings of the three aforementioned fields.

2.1 Differential Privacy

Differential privacy is a mathematical framework for quantifying the anonymization of sensitive data, and it has shown its strong capacity in privacy guarantees for aggregation on datasets and databases. In this thesis, we use differential privacy as the privacy-preserving mechanism to protect the privacy of raw datasets stored in each local worker node.

As a general term, a query function f is used as the mapping from databases to real entries, and the true answer is the consequence of applying f to the database. In order to protect the true answers from being recognized by attackers, the values returned to the users are the true answer plus random noise generated based on a specific distribution.

Until 2005, most of the works in privacy protection focused on using noisy sums. Blum termed the query with (slightly) noisy reply as Sub-Linear Queries (SuLQ) in (19). As stated in the paper, the query function is $f = \sum_i g(x_i)$, in which x_i represents the i^{th} row of a statistical database and g maps rows in the database to {0,1}. The paper defined that a database query mechanism is (ϵ, δ, T)-private when the following

formula is valid for every data element of index i, for every predicate $f : D \to \{0, 1\}$ in which element d_i is drawn from an arbitrary domain D, and for every adversary making at most T queries, given that the data is extracted from a distribution with enough generality but without missing much information.

$$Pr\left[conf(p_T^{i,f}) - conf(p_0^{i,f}) > \epsilon\right] \leq \delta$$

Here (19) assumes that rows of the database are independent. And for any predicate $f : D \to \{0, 1\}$, $p_0^{i,f}$ denotes a priori belief that $f(d_i) = 1$ and $p_T^{i,f}$ denotes a posteriori belief that $f(d_i) = 1$ giving the answers to T queries. As a result, a randomly generated noise of $N(0, R)$ where $R = R(\epsilon, \delta, T)$ is added to the true answer $\sum_{i \in S} f(d_i)$ where S is a set of rows in the database within mapping $g(x_i)$. We can see the power of the noisy sums query scheme based on the advanced examples of carrying out standard data mining tasks using SuLQ, such as Principal component analysis (PCA) in dimension reduction, k-means clustering, ID3 classification, and statistical queries learning model. However, a more sophisticated privacy preserving scheme is required when the machine learning task becomes much more complicated.

The term differential privacy was introduced by Dwork as a privacy-preserving mechanism for statistical databases in a series of studies (20) (14) (21) (22). Within the initial concept of differential privacy, it protects data privacy by adding random noise to the real entries in the database, in which the noise is generated according to a discreetly selected distribution.

As mentioned in (20), the ultimate goal of a privacy-preserving statistical database is to empower every user to learn the properties of the whole population but still protect the privacy of every individual data owner. In 2006, Dwork proposed that when considering f as the mapping from database to vectors of real entries, they can prove that the data privacy can be preserved by calibrating the standard deviation of the noise adding mechanism based on the sensitivity of the function f. In this paper, they model a database as a vector of n entries from some domain D, and consider domain of the form $(\{0, 1\})^d$ or R^d. The sensitivity $S(f)$ is defined as an inherent quantity in f, which is independent of the database. L_1 sensitivity of a function $f : D^n \to R^d$ is defined as the smallest number $S(f)$ such that for all $x, x' \in D^n$ which differ in a single entry,

$$\left\|f(\mathbf{x}) - f\left(\mathbf{x}'\right)\right\|_1 \leq S(f)$$

. According to the theory, if noise Y is a vector of d independent Laplace variables, the density function at y will be proportional to $exp(-\|y\|_1/\lambda)$. As a consequence, the random variables $z + Y$ and $z' + Y$ will be as follows for all $t \in R^d$

$$\frac{Pr(z + Y = t)}{Pr\left(z' + Y = t\right)} \in \exp\left(\pm\frac{\|z - z'\|_1}{\lambda}\right)$$

. Thus, it is sufficient to add Laplace noise with standard deviation of $S(f)/\epsilon$ in each coordinate to make sure the returned value is with sufficient privacy.

Moreover, a noise adding mechanism is defined as ϵ-indistinguishable in (20) if for all pairs $x, x' \in D^n$ which differ in only one single entry, for all adversaries $\mathcal{A}$, and for all transcripts t:

$$\left|\ln\left(\frac{Pr\left[\mathcal{T}_{\mathcal{A}}(\mathbf{x}) = t\right]}{Pr\left[\mathcal{T}_{\mathcal{A}}\left(\mathbf{x}'\right) = t\right]}\right)\right| \leq \epsilon$$

. Sometimes when ϵ is small, $ln(1 + \epsilon) \approx \epsilon$, and the ϵ-indistinguishable definition will be approximate equivalent to $\frac{Pr[\mathcal{T}_{\mathcal{A}}(\mathbf{x})=t]}{Pr[\mathcal{T}_{\mathcal{A}}(\mathbf{x}')=t]} \in 1 \pm \epsilon$.

In other words, as stated in (14), the definition of ϵ-indistinguishability can be defined as follows for two datasets that only differ on one row, if the respective output random variables of query responses τ and τ' satisfy for all sets S of responses :

$$Pr[\tau \in S] \leq \exp(\epsilon) \times Pr\left[\tau' \in S\right]$$

. Similarly, a noise adding mechanism can be defined as δ-approximate ϵ-indistinguishable under the same conditions that:

$$Pr[\tau \in S] \leq \exp(\epsilon) \times Pr\left[\tau' \in S\right] + \delta$$

. The non-zero δ allows people to release the strict relative shift when events are not especially likely to happen.

In the following 10 years, during the rapid development of machine learning techniques with neural networks, privacy-utility issue has become the core when training complex models with large-scale crowdsourced dataset containing sensitive information. In 2016, Google developed a new algorithmic technique to help developers and researchers better deal with the privacy issue. They implemented differential privacy within their machine learning framework TensorFlow in Python, which is the differentially private stochastic gradient descent algorithm (DPSGD) (13).

Within DPSGD, they take the input of: (1) data samples $\{x_1, ..., x_N\}$, (2) loss function $L(\theta) = \frac{1}{N}\sum_i L(\theta, x_i)$ in which θ denotes the parameters of the model, and (3) parameters of learning rate η_t, noise scale σ, group size L, and gradient norm bound C. After randomly initiating θ_0, they do the following steps for every $t \in T$: (1) take a random sample L_t with the sampling probability of L/N, (2) compute gradient for each $i \in L_t$ by calculating $g_t(x_i) \leftarrow \bigtriangledown_{\theta_t} L(\theta_t, x_i)$, (3) clip gradient by calculating $\overline{\mathbf{g}}_t(x_i) \leftarrow \mathbf{g}_t(x_i) / \max\left(1, \frac{\|\mathbf{g}_t(x_i)\|_2}{C}\right)$, (4) add noise by calculating $\tilde{\mathbf{g}}_t \leftarrow \frac{1}{L}\left(\sum_i \overline{\mathbf{g}}_t(x_i) + \mathcal{N}\left(0, \sigma^2 C^2 \mathbf{I}\right)\right)$, and (5) perform descent by calculating $\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{\mathbf{g}}_t$. The output are the parameters θ_T and the overall privacy cost (ϵ, δ) calculated using a privacy accounting method.

Considering the federated learning framework in this thesis is implemented with PyTorch, we use Opacus, the PyTorch version implementation of DPSGD realeased by Facebook, to handle the differential privacy part of the experiments.

2.2 Federated Learning

2.2.1 Definition of Federated Learning

The term federated learning was first defined in 2016 by McMahan from Google (3) as "the learning task is solved by a loose federated of participating devices (which we refer to as clients) which are coordinated by a central server". On account of the development of modern mobile devices, the massive amount of data obtained by smart devices is of great value and importance to machine learning tasks such as text recognition from audio, image detection from pictures, and disease prediction based on personal medical records. However, this rich data usually comes with high sensitivity which prevents uploading all data into a data center and train a model there. Thus, federated learning framework is proposed to leave the training data on the local mobile devices and learn a shared model by aggregating the locally computed updates. Within the framework of federated learning, clients have their own local training dataset and this dataset will never be uploaded to the central server. The only thing that each client uploads to and be broadcasted from the central server is the parameters of the model.

Aside form the definition of federated learning mentioned above, (3) also introduced an algorithm called Federated Averaging (FedAvg), combining local stochastic gradient descent (SGD) on each client with a central server which performs averaging on model

parameters. At that time, they also pointed out that the unbalanced and non-identically and independently distributed (non-IID) data partitioning among a huge number of unreliable devices together with the limited communication bandwidth were the largest challenges of federated learning.

2.2.2 Cross-silo and Cross-device Federated Learning

Later in 2019, two types of federated learning settings were introduced, specifically the "cross-device" federated learning setting and "cross-silo" federated learning setting. The major difference of these two definitions can be found in Table 2.1.

Features	Cross-silo federated learning setting	Cross-device federated learning setting
Client type	A small number of reliable clients (e.g. 2-100 different medical or financial organizations).	A large number of unreliable clients (e.g. 10-million mobile devices or IoT sensors).
Data partition	Both possible for example-partitioning (horizontal) and feature-partitioning (vertical).	Only possible for example-partitioning (horizontal).
Addressability	Each client has its own unique identity.	No client identifiers in use.
Client statefulness	Stateful (each client can participate in each round of the computation and carry state from previous states).	Stateless (each client is more likely to participate only once in a task).

Table 2.1: Major difference between cross-silo and cross-device federated learning setting.

With these two types of setting, (4) proposed a broader definition of federated learning as "a machine learning setting where multiple entities (clients) collaborate in solving a machine learning problem, under the coordination of a central server or service provider. Each client's raw data is stored locally and not exchanged or transferred; instead, focused updates intended for immediate aggregation are used to achieve the learning objective." According to the broader definition, the focused updates narrows the scope of information being communicated between the central server and clients to only contains the minimal information which is essential for the learning task. Also, the aggregation on central server will be performed as early as possible to guarantee the data minimization.

Considering the cross-silo federated learning, there are plenty of applications in various fields including financial fraud detection, medical image classification for pre-diagnosis, and smart manufacturing. A detailed example of cross-silo federated learning setting in financial field is the collaboration between WeBank and Swiss Re (23). WeBank is an online lender owned by Chinese high-tech company Tencent, and the AI team within WeBank has created "Federated AI Technology Enabler (FATE)", an industrial-level open-source technical framework. And Swiss Re is a leading company in the field of reinsurance. After signing the partnership, these two large organizations would study the efficiency challenges imposed by data silos with the help of federated

learning (24). An example of cross-silo federated learning setting in medical field is the MesoNet (25). It uses a deep convolutional neural networks (CNN) model to predict the overall survival of mesothelioma patients. The CNN model MesoNet is validated on dataset from MESOBANK, which is an international organization provides collection of high-quality samples from mesothelioma patients around EU. Since MESOBANK offers dataset to medical researchers from EEA, USA, Canada, Australia and New Zealand, data retrieved from MESOBANK are preprocessed by adding noise on under a federated learning setting.

On the other hand, the cross-device federated learning has been widely deployed in the field of digital customer analysis. For example, Google has introduced federated learning in their research of mobile keyboard prediction (26) (27) (28) (29). Google trained a recurrent neural network (RNN) language model in the federated learning setting for the next-work prediction of virtual keyboard on smartphones (Google Keyboard, also called Gboard). Compared with server-based training, the federated learning setting with FedAvg algorithm has shown a better performance (precision recall) and guaranteed the privacy of user sensitive data at the same time (26). Other than normal characters from more than 600 languages, the RNN model can also predict emoji from the previous typed text on Gboard. Its great performance shows the feasibility of implementing federated learning setting as the framework to train a production-quality model in the field of natural language understanding and keep user data locally to avoid interfering user data privacy (28). The federated learning setting with the character-level RNN model has also demonstrated its capability to learn out-of-vocabulary (OOV) words as stated in (29). Aside from the regular next-word prediction task, Google also introduced federated learning setting to improve the quality of search suggestion based on the data from more than 1 billion end users of Gboard (27).

Considering the scope of this thesis, we use the cross-silo federated learning setting to study the effect of data bias in differentially private federate learning among a small number of reliable entities.

2.2.3 Horizontal and Vertical Federated Learning

Aside from classifying federated learning settings based on their clients, another important concept within federated learning is the feature space and sample ID space according to the data distribution characteristics among clients. Based on the difference on

intersection of feature space and sample space, horizontal federated learning, vertical federated learning, and federated transfer learning are proposed in (30). Horizontal federated learning is also called sample-based federated learning since the datasets are sharing the same feature space but different sample IDs under this scenario. For example, the business of several regional banks could be quite similar but their user group could be largely different. Thus, these banks are sharing the same feature space but different sample space. On the contrary, vertical federated learning can be named as feature-based federated learning, in which several datasets share different feature space but same sample ID space. A real-world use case can be seen in the e-commerce field. A bank and an e-commerce company in the same region might share the same group of users, but they definitely do not have the same features in their dataset. However, the user browsing history and purchasing transactions from the e-commerce company and the revenue with expenditure behavior records from the bank can be collaboratively used to predict the next product purchase behavior of a user. In this task, the bank and the e-commerce company have the same sample space but different feature space. In between horizontal and vertical federated learning, federated transfer learning depicts the scenarios that several datasets differ in both sample space and feature space, leading to a relatively small intersection among entities. Considering one entity as a Chinese bank and another entity as an American e-commerce company. In this scenario, both sample space and feature have quite small intersection. Thus, a common representation is learned based on the limited number of common samples.

Considering the scope of this thesis, we use the horizontal federated learning setting to study the effect of data bias in differentially private federate learning, in the situation that different local nodes share the same feature space but own samples with different sample IDs.

2.2.4 Significant Challenges in Federated Learning

As introduced above, federated learning is a convincing approach for collaboratively training a model but keep the data decentralized in local clients. However, the model performance in federated learning is not only influenced by the federated algorithm, but also largely impacted by the data distribution among clients (4). Two of the most significant challenges in federated learning are class imbalance and non-IID data distribution

data. In this paper, we summarize these two challenges as data bias and perform systematically investigation on the effect of data bias in federated learning with differential privacy.

2.3 Data Imbalance

2.3.1 Data Imbalance in Classification

In a machine learning problem with classification task, data imbalance refers to the the uneven number of samples of each class or of each label.

In binary classification tasks, data imbalance is a notable issue especially in cases where samples with negative label largely outnumber samples with positive label, including computer-assisted medical diagnostics using images and test reports, network attack detection, unreliable telecommunication customer detection, and financial fraud detection (31).

Common methods dealing with data imbalance of classification tasks in single machine (without distributed setting) could be categorized as data driven approaches and algorithm driven approaches (32). Data driven approaches include: (1) under-sampling by discarding samples from majority class to make an even number of samples between classes such as random under sampling (33), and (2) oversampling by replicating samples from the minority class such as synthetic minority over-sampling technique (SMOTE) (34). Algorithm driven approaches include: (1) cost-sensitive learning by choosing a class with minimal conditional risk in order to minimize misclassification cost such as MetaCost (35), and (2) thresholding by adjusting the decision threshold of a sample according to the class to reduce misclassification cost and improve performance at the same time (36), and (3) hybrid methods by combining ensemble learning and sampling to increase the accuracy of minority class while keeping the accuracy of the overall dataset in a reasonable range such as SMOTEBoost (37).

When there is a class imbalance in the dataset, the proportion and number of samples of minority classes are significantly lower than the majority classes. In some cases, the minority classes are undoubtedly playing more important role compared with their proportion in the whole dataset. For example, the resulting reduction of classification accuracy on minority classes could lead to a bad consequence when doing a sudden disease prediction based on abnormal heart rates.

With the context of this thesis, the data imbalance of classification task is measured from perspectives of both class imbalance and label imbalance.

2.3.2 Data Imbalance in Federated Learning

When the dataset is distributed among several clients rather than being trained in a single node, the impact of data imbalance is even worse. For example, in the series of federated learning framework with next-word prediction task of Gboard from Google, the SOS typing is quite uncommon among all device (smartphones in this case), but SOS is much more important than any of the big name restaurants. The minority class SOS even needs higher prediction accuracy than other majority classes (26).

Considering the strict constraint of communication content in the setting of federated learning, it is not possible to upload additional information needed for conventional methods to mitigate the impact of data imbalance (38) (39) (40). Thus, methods mentioned above are not suitable for federated learning data imbalance issue since neither the clients nor the central server has the full access to the complete training set (41).

In real-world machine learning tasks, the class imbalance phenomenon happens quite often. For example, the number of patients diagnosed with different diseases can vary significantly in medical image classification tasks (42) (43). Also, the Google research of predicting emoji with Gboard also pointed out that people could have largely different personal preferences in their daily use of the virtual keyboard on smartphones (29).

Some studies have proposed methods to be added in local clients in order to solve the data imbalance issue in federated learning without uploading additional data distribution information to the central server. Inspired by the observation that minority classes usually contains very few instances with relatively high degree of visual variability, (44) proposed to learn a Euclidean embedding $f(x)$ from an image x to the feature space R^d, so that the embedded features are discriminative without local class imbalance. With the help of transfer learning, (45) proposed that the knowledge from data-rich classes in the head of the distribution can be encoded with a meta-network and then be gradually transferred from head to body and from body to tail. However, these methods would not work when there is a mismatch between the local data distribution and the global data distribution. They might even result in a negative side-effect on the model in central server within a federated learning setting.

(41) defined the class imbalance as two classes, the local imbalance and the global imbalance. The local class imbalance γ_j of client j is defined as the ratio between number of samples of majority class and minority class, in which N_p^j denotes the number of samples in class p on client j:

$$\gamma_j = max_p\{N_p^j\}/min_p\{N_p^j\}$$

. And the global class imbalance Γ is defined as the ratio between total number of sampler of majority class and minority class:

$$\Gamma_j = max_p\{\sum_j N_p^j\}/min_p\{\sum_j N_p^j\}$$

. In order to quantify the mismatch between local class imbalance and global class imbalance, they use Q to represent the overall number of classes, use vector $v_j = [N_1^j, ..., N_Q^j]$ to denote the local data composition on client j, and use vector $V = [\sum_j N_1^j, ..., \sum_j N_Q^j]$ to denote the global data composition. After that, cosine similarity (CS) score is used to measure the similarity between two data compositions as:

$$CS_j = (v_j \cdot V)/(\|v_j\| \, \|V\|)$$

.

Aside from quantifying global class imbalance and local class imbalance, they proposed a monitoring scheme which can estimate the composition of training data across classes during each federated training round. This is designed to alert administrator when certain imbalanced data composition appears. This paper also designed the Ratio Loss function to mitigate the impact of class imbalance in federated learning.

With the context of this thesis, we use a highly imbalanced Adult dataset to introduce the data imbalance into the federated learning framework, and create a comprehensive mechanism to measure the effect of data imbalance by calculating the model performance of subgroups in the whole dataset.

2.3.3 Data Imbalance in Differential Privacy

Moreover, there are also studies show that data imbalance also has great influence on differentially privacy training mechanisms.

As stated in (46), differential privacy might also deteriorate the existing data bias in the raw dataset and results in a largely different accuracy on different subgroups of the dataset. This paper carried out experiments of DP-SGD in a binary classification task on a single machine. The range of imbalance was set from 0.1% to 30%, and the privacy budget ϵ was set from 1.15 to 16.2. Their experiment results demonstrated that an increasing data imbalance trained with differential privacy mechanism leads to a significantly increasing disparity of accuracy between 2 classes (subgroups). Thus, applying differential privacy on dataset with imbalance will result in a huge impact on the model performance of minority subgroups even with loose guarantees.

In this thesis, we also compare the model performance of subgroups in different differential privacy settings, so that we can analyze what impact the privacy-preserving mechanism would introduce when the machine learning task is executed on a highly imbalanced dataset.

2.4 Non-IID Data Distribution in Federated Learning

As we mentioned in chapter 1, we define data bias from two perspectives, one among all clients in the federated learning setting, another one within a specific client. Since the previous section has fully introduced data bias within one specific machine, we will introduce the data bias from the among-clients perspective in this section as the data distribution scenarios.

In real-world machine learning tasks, datasets owned by several clients often comes with various data formats and some unique preferences based on their business background. Thus, the resulting diversity would slow down the convergence of the global model in a federated learning setting.

In the experiments of paper (47), they assigned every client exact m classes of the dataset to mimic the "non-IID (m)" federated learning setting. Their results showed that models with IID data distribution has a relatively faster convergence than non-IID ones.

A more comprehensive experimental study on the federated learning with non-IID data silos showed that nowadays there is no single federated learning algorithm that could outperform others considering all possible scenarios of the non-IID data distribution (48). This paper considered 3 types of possible non-IID data distribution cases,

specifically (1) label distribution skew, (2) feature distribution skew, and (3) quantity skew. The 6 data partitioning strategies can be seen in Table 2.2.

Distribution type	Imbalance type
Label distribution skew	Quantity-based label imbalance
Label distribution skew	Distribution-based label imbalance
Feature distribution skew	Noise-based feature imbalance
Feature distribution skew	Synthetic feature imbalance
Feature distribution skew	Real-world feature imbalance
Quantity skew	

Table 2.2: 6 data partitioning strategies

In label distribution skew, a simple case in practice is that some hospitals have great specialization in a particular set of disease categories, then the number of patient records of these diseases within these hospitals would naturally be much higher. The difference between quantity-based and distribution-based label imbalance is the number of labels in each client. Quantitiy-based label imbalance assigns a fixed number of labels in a client, and is considered as an extreme case. Each client in distribution-based label imbalance is allocated a certain proportion of the samples within each label according to the Dirichlet distribution, which is normally used as the prior distribution in Bayesian statistics and as the appropriate method to simulate real-world data distribution scenarios (49).

In feature distribution skew, examples could be different fur colors and various patterns in different areas for cats. In the noise-based feature imbalance, different levels of Gussian noise is added to the randomly and equally divided parts of the whole dataset to simulate different noise level among all clients. In the synthetic feature imbalance, they divided a cube into 8 equal-volume parts and allocated two symmetric parts in one client, creating a label-balanced but varied feature data distribution. The real-world feature imbalance uses the inherent feature of data samples to distribute, such as distributing the EMNIST dataset of handwritten characters/digits based on their writers.

In quantity skew, the number of samples within each local dataset are different among all clients. Dirichlet distribution is also used here to allocate different number of samples to each client.

In the context of this thesis, we consider the data partition strategies from the label and class perspectives. Chapter 3 introduces the 4 data distribution scenarios in details and indicates how we use these scenarios to mimic real-world maching learning task settings.

3

Data Bias

As mentioned in Chapter 1, this thesis measures data bias from the following two aspects: (1) data imbalance within the dataset, and (2) data distribution among entities. This chapter first introduces the highly imbalanced Adult dataset to show the label imbalance and target class imbalance within a dataset. Next, this chapter shows 4 representative data distribution schemes in detail to illustrate how to mimic data distribution scenarios in real-world machine learning tasks. With the detailed illustration of both data imbalance and data distribution scenarios, the data bias is thoroughly defined and chapter 4 can then present the experiment design to systematically measure the effect of data bias.

3.1 Data Imbalance

3.1.1 Adult Dataset

The most famous datasets in machine learning classification tasks are MNIST (15) and CIFAR (50). These two image classification dataset are great benchmark for balanced dataset since they both contain almost even number of sampler per class. However, a great balance in the number of samples per class is not suitable when we want to investigate the effect of data bias. Thus, we choose to use the Adult dataset, which is a multivariate dataset derived from real world, with binary classification task and several categorical features.

The adult dataset was extracted from the 1994 Census bureau database, and its task is to classify whether a given adult makes more than $50,000 a year (51).

The target class is salary in the adult dataset, where every sample's annual income is being classified as either >50K or <=50K. Features of the adult dataset include age, workclass, education level, marital status, occupation, race, sex, capital status, hours per week, and nationality.

In order to distribute the adult dataset in a federated learning setting and investigate whether there is difference on performance between minority samples and majority samples, we made a Sex_Race label by combining Sex feature values and Race feature values for each sample in the dataset. The Sex_Race label has 10 classes in total given that the Sex feature is consisted of Female and Male and the Race feature includes Amer-Indian-Eskimo, Asian-Pac-Islander, Black, Other, and White.

In the following sections of this thesis, dataset refers to the Adult dataset, in which salary as target class and Sex_Race as label, unless otherwise noted.

3.1.2 Data Imbalance within Adult Dataset

The complete adult dataset had 46033 samples, in which 34611 samples are in the target class of <=50K and 11422 samples are in the target class of >50K. As a result, the <=50K class has more than three times of the number of samples of >50K class.

Figure 3.1 shows the number of samples per Sex_Race label of the whole dataset, and the exact number of samples per Sex_Race label per target class is shown in table 3.1. It is clear that there is a huge difference on the number of samples between the majority groups and minority groups. As stated in table 3.1, the Male_White label contains 27421 samples, and it is almost 60% of the complete dataset. And the Female_White label contains 12023 samples, which is more than 25% if the complete dataset. On the contrary, the Female_Other label only has 135 samples, and this is less than 0.3% of the complete dataset. The Female_Amer-Indian-Eskimo label only has 166 samples, which is less than 0.4% of the complete dataset.

As a conclusion, the adult dataset is a highly imbalanced dataset considering from perspectives of the target class and the Sex_Race label. We hereby define "<=50K" as the majority class and ">50K" as the minority class in adult dataset. Also, samples of Sex_Race label "Female_Amer-Indian-Eskimo" and "Female_Other" are defined as under-represented groups within the adult dataset. Likely, samples of Sex_Race label "Female_White" and "Male_White" are defined as over-represented groups within the adult dataset.

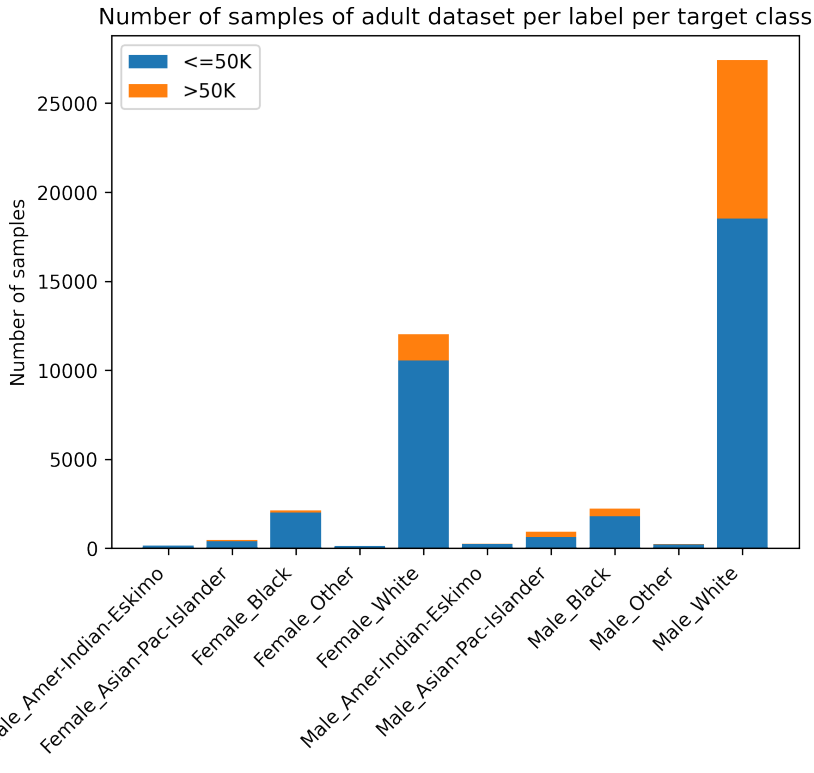

Figure 3.1: Number of samples per target class per Sex_Race label in complete adult dataset

3.1.3 80/20 Train-Test Split of Adult Dataset

Normally we split the whole dataset based on its target class. However, after introducing Sex_Race label to the Adult dataset, we also need to adjust the train-test split scheme. We now split the complete adult dataset as train set and test set based on both Salary target class and Sex_Race label, using a 80/20 proportion. This is to guarantee that the testing results would be able to show the model performance on the complete dataset, per target class, and per label. Table 3.2 shows the exact number of samples per Salary target class after the dataset split. And table 3.3 gives the number of samples per Sex_Race label on train set and test set. Figure 3.2 and figure 3.3 illustrate the number of samples per target per class in the train set and test set.

Sex_Race label & Target class	<=50K	>50K	total	total %
Female_Amer-Indian-Eskimo	152	14	166	0.3606 %
Female_Asian-Pac-Islander	401	69	470	1.0210 %
Female_Black	1998	127	2125	4.6163 %
Female_Other	126	9	135	0.2933 %
Female_White	12023	10542	1475	26.1182 %
Male_Amer-Indian-Eskimo	230	39	269	0.35844 %
Male_Asian-Pac-Islander	619	334	953	2.0703 %
Male_Black	1806	425	2231	4.8465 %
Male_Other	202	38	240	0.5214 %
Male_White	18529	8892	27421	59.5681 %
total	34611	11422	46033	100 %

Table 3.1: Number of samples per target class per Sex_Race label in complete adult dataset

	train set	test set
<=50K	27684	6927
>50K	9135	2287

Table 3.2: Number of samples per Salary target class on train set and test set

	train set	test set
Female_Amer-Indian-Eskimo	132	34
Female_Asian-Pac-Islander	375	95
Female_Black	1699	426
Female_Other	107	28
Female_White	9618	2405
Male_Amer-Indian-Eskimo	215	54
Male_Asian-Pac-Islander	762	191
Male_Black	1784	447
Male_Other	191	49
Male_White	21936	5485

Table 3.3: Number of samples per Sex_Race label on train set and test set

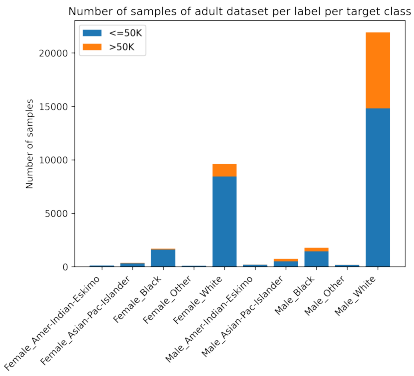

Figure 3.2: Number of samples per target class per Sex_Race label in adult train set

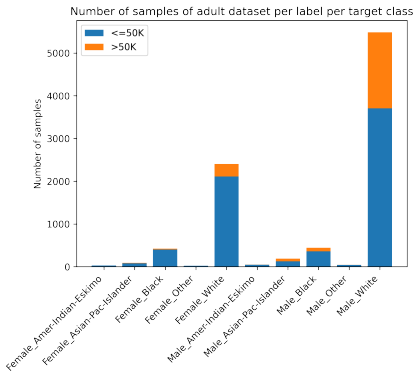

Figure 3.3: Number of samples per target class per Sex_Race label in adult test set

3.2 Data Distribution Scenarios

We will illustrate how to simulate the various data distribution schemes under a distributed machine learning setting. The implementation as the data distribution module within a federated learning framework will be released later on.

As mentioned in chapter 1, the real-world organizations in a distributed machine learning setting are usually with various data composition, which lead to the non-IID data distribution among all worker nodes. Thus, in order to measure the effect of data bias, we choose to simulate 4 types of representative data distributions in this thesis, which are fully IID data distribution, fully non-IID data distribution, partial IID data distribution, and statistical distribution.

The real-world machine learning problems have a relatively lower possibility to have the same number of samples of every target class on every worker node, so we do not distribute the dataset based on the target class. On the other hand, the label of each sample is easier to be obtained. For example, the collection date and time of the sensor data itself could be its label, and the demographic attributes like sex, rage, and age could also be regarded as the label of the user data. Thus, we distribute the dataset among worker nodes based on their label, which will also be the scheme in our experiments.

3.2.1 Fully IID Data Distribution

As describe by its name, fully IID data distribution means that every worker node would have the same number of samples on each label, inherently creates the same number of samples in total on every worker node.

Fully IID data distribution is the basic assumption of many studies in distributed machine learning setting. It is very helpful when creating the benchmark for a specific machine learning model under collaborative training. However, the complexity of real-world machine learning problems makes the fully IID data distribution rare. Thus, this thesis regards the fully IID data distribution as an ideal baseline and puts more attention on other data distribution scenarios.

3.2.2 Fully N-class Non-IID Data Distribution

On the opposite of having the same number of samples per label on each worker node, fully non-IID data distribution shows us the extreme data distribution scenarios. In this paper, we use n-class scheme for the non-IID data distribution scenarios.

The n-class non-IID data distribution means that the number of unique labels in a worker node is exact n. To achieve this objective, we need to partition all samples within one label as multiple parts, and arrange the labels in multiple worker nodes. Details can be found in `data.py` within the `Code` folder.

The following formula shows our calculation method for the number of partitions of samples with the same label.

$$number\ of\ partitions\ within\ one\ label = \frac{number\ of\ worker\ nodes\ *\ n}{number\ of\ labels\ in\ complete\ dataset}$$

Here we take an example to show the fully n-class data distribution scenario. Assume that the dataset has 10 labels, and we want to make a 2-class fully non-IID data distribution of this dataset among 20 worker nodes. Then the number of partitions we need for each label is $\frac{20*2}{10} = 4$. Since there is no remainder in this formula, this is a valid n-class fully non-IID data distribution scheme. This formula means that we need to partition the samples of each label as 4 parts, and the arrange which labels should be distributed in which worker nodes.

With the formula above, we can do a validity check before really distribute the dataset, in order to avoid wasting computing resources. There are several invalid situations in a n-class fully non-IID data distribution, which are listed as below:

- There are labels left which are not allocated to any one of the worker nodes.
  For example, a 2-class fully non-IID scheme on a 10-label dataset with 3 worker nodes is invalid since $\frac{3*2}{10} = 0$ ... 6.

- There are different number of partitions within each label.
  For example, a 5-class fully non-IID scheme on a 10-label dataset with 3 worker nodes is invalid since $\frac{3*5}{10} = 1$ ... 5. The remainder value of 5 means that the number of partitions among all labels are not the same, which is not acceptable in our n-class fully non-IID scenario.

3.2.3 Partial N-class Non-IID Data Distribution

In between the extremely orderly fully IID data distribution and the extremely uneven n-class fully non-IID data ditribution, there are partial IID data distributions. The proportion of non-IID part in the complete dataset should be defined before actually distribute the data. This parameter could be any number between 0 and 100, including 0 and 100 but not limited to integers. During our experiments on the adult dataset, we will pick up a series of non-IID proportion to investigate the effect of different data distribution with the gradually increase non-IID percentage.

The IID part of the complete dataset will be distributed using the fully IID scheme, and the non-IID part of the complete dataset will be distributed based on the fully n-class non-IID scheme. Thus, we can also do a validity check before the calculation to ensure it is a valid setting for partial n-class non-IID data distribution.

For example, we want to simulate a 30pct 2-class non-IID data distribution with 10 labels in dataset and 10 worker nodes. The validity check is passed according to $\frac{10*2}{10} = 2$. The complete dataset will be split as 30% and 70% first. Then apply fully 2-class non-IID data simulation scheme on the first partition of 30%, and distribute the second partition of 70% under the fully IID data distribution scheme.

3.2.4 Statistical distribution

Some of the real-world classification problems can also be in a statistical distribution, such as binomial distribution for flipping a coin for thousands of times and normal distribution for the score of every student in a large course. Considering the real use cases, we choose to simulate the normal distribution in our experiments. We need to

set the μ and σ for the normal distribution of each label, and these parameters can be all different or all the same among all labels. After having the $N(\mu, \sigma^2)$ for each label, we need to run a random simulation to generate as many data points as the number of samples in this label. Then we equally divide the x-axis as k parts, in which k is the number of worker nodes in the federated setting. As a result, samples within one specific label will be distributed among worker nodes according to the number of randomly generated normal distribution samples that fall in its interval.

For example, we can set $N(0, 1)$ for half of the labels and $N(1, 4)$ for another half of the labels, and distribute the samples based on the amount of generated data points fall in a worker node's interval.

The exact experiment setting of theses 4 types of data distribution scenarios will be illustrated in detail in the next chapter.

4

Experiment Design

This chapter describes the methodology of measuring the effect of data bias in differentially private federated learning. In particular, we propose a comprehensive experiment scheme considering three dimensions: (1) variation on data distribution scenario, (2) different privacy budget of differential privacy, and (3) model parameter fusion scheme in distributed training setting. The 6 data distribution scenarios and 10 differential privacy setting will be thoroughly explained in this chapter. Also, this chapter illustrates the detailed experiment setting including the machine learning problem and the federated learning setup.

4.1 Measurement of Utility and Fairness

As introduced in chapter 1, this thesis investigates the effect of data bias on differentially private federated learning, by measuring the impact of privacy and data distribution mechanism on the utility and fairness of a distributed machine learning system. To achieve this objective in a classification task, we need to measure the utility using a series of metrics derived from the confusion matrix, including accuracy, precision, recall, and F_1 score. Moreover, in order to measure the fairness of the classification model, we also need to compare the overall model performance with the per-class and per-label model performance to see if there is a large difference between the general utility and the partial utility.

The details of metrics and testing scheme will be introduced in the following sections.

Aside from vertically measuring the utility and fairness within one specific experiment, we also need to respectively conduct comparison among experiments to examine the effect of various data distribution scenarios and different privacy levels. To be more specific, the first horizontal comparison will be in a fixed data distribution scenario but with different privacy levels. The second horizontal comparison is conducted with the fixed differential privacy budget among different data distribution scenarios.

4.2 Basic Experiment Setting

4.2.1 Machine Learning Problem

Dataset As mentioned in chapter 3, data imbalance is quite normal in real-world machine learning problems, so we choose the Adult dataset with Sex_Race label to be our dataset for all experiments regarding a classification task. The target class in adult dataset is Salary, and samples will be classified as "<=50K" or ">50K" for the annual income. Besides target class, we created Sex_Race label to help us better simulate different data distribution scenarios. Values in Sex_Race label of adult dataset are: 'Female_Amer-Indian-Eskimo', 'Female_Asian-Pac-Islander', 'Female_Black', 'Female_Other', 'Female_White', 'Male_Amer-Indian-Eskimo', 'Male_Asian-Pac-Islander', 'Male_Black', 'Male_Other', 'Male_White'. We use this adult dataset with Sex_Race label in all of our following experiment design, result, and analysis.

Task The task is to perform a binary classification on the Salary. Since this thesis pays most of the attention on the data distribution scenarios and differential privacy budget, we decided to use a relatively simple Neural Network (NN) classification model to help us avoid the influence from the machine learning model itself.

Model The NN classification model is consisted of 3 fully connected linear layer. Table 4.1 shows the detailed setting of each layer in the NN model. Figure 4.1 visualizes the gradients of the simple NN classification model. This figure illustrates the pytorch operations of the model with the help of package torchviz (52). And the figure is built during forward propagation and shows which operations can be called on backward propagation.

Layer name	Layer type	Number of in features	Number of out features	Bias or not
fc1	linear	38	32	True
fc2	linear	32	32	True
fc3	linear	32	2	True

Table 4.1: Details of each layer in the NN binary classification model

4.2.2 Federated Learning Setup

In all of our experiments, we use 11 nodes in total, including 10 worker nodes used for training and 1 central server being responsible for orchestration. In the training process, we use 500 rounds with 1 epoch , in order to guarantee the convergence of the simple NN binary classification model. Also, we use Federated Averaging (FedAvg) as the fusion scheme in our federated learning framework. The main idea in FedAvg is that the central server would boradcast the averaged value of each worker node for every model parameter. With this fusion scheme, the parameter values in the central server node is a good representative of all worker nodes.

4.3 Data Distribution Scenarios

To show the effect of different data distribution with a certain level of continuity, we choose 3 non-IID proportions in the partial n-class non-IID data distribution scenario. With that, we can see the impact that an increasing proportion of non-IID part in the dataset has on the overall, per-class, and per-label performance.

In our experiment, we simulate the following 6 data distribution scenarios: (1) fully IID, (2) 30% 2-class non-IID, (3) 50% 2-class non-IID, (4) 70% 2-class non-IID, (5) fully 2-class non-IID, (6) normal distribution.

4.3.1 Scenario 1: Fully IID data distribution

Figure 4.2 shows the fully IID data distribution scenario of adult dataset among 10 worker nodes. In this scenario, each worker node has the same number of samples in total and per label.

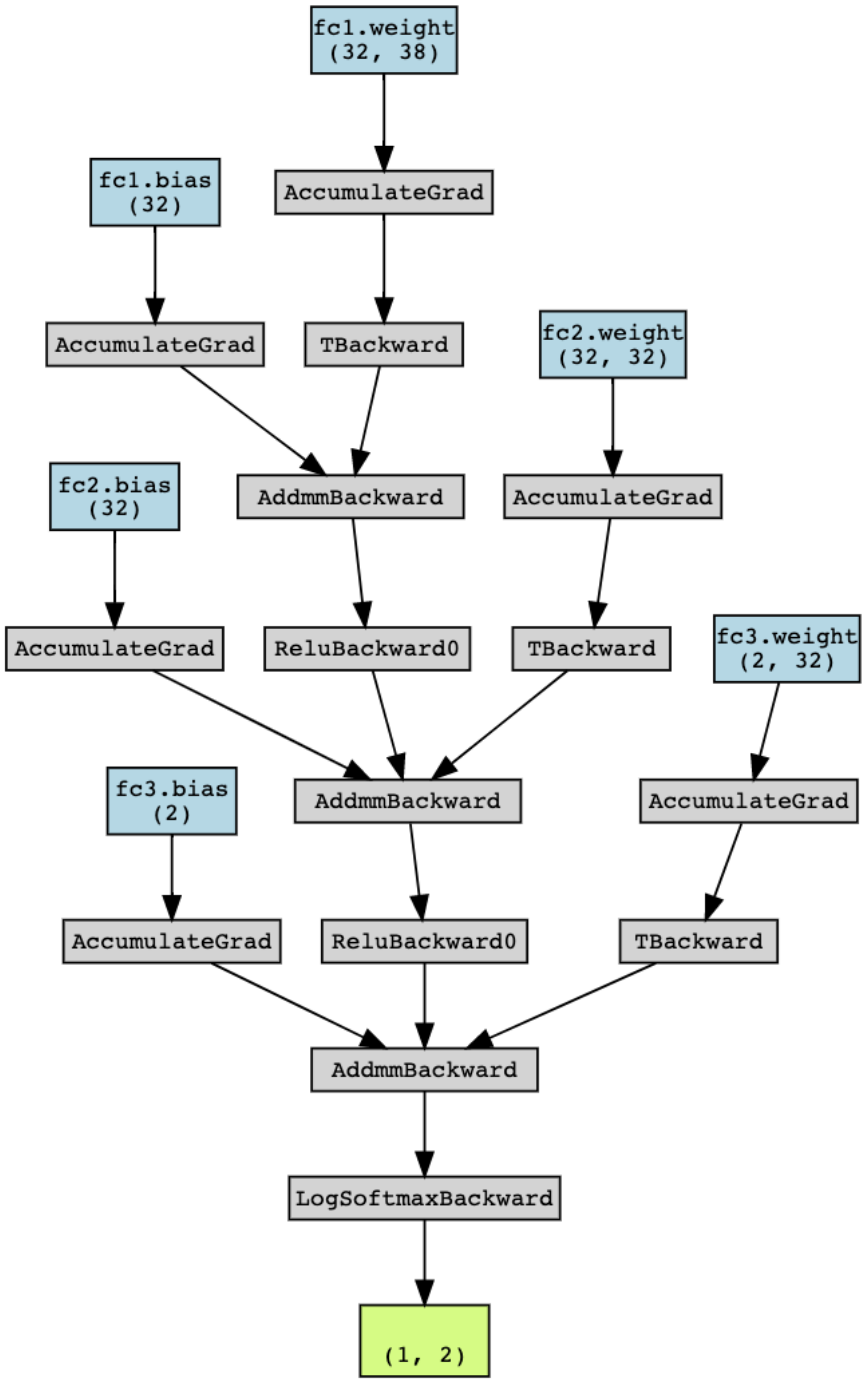

Figure 4.1: Visualization of binary classification model for adult dataset

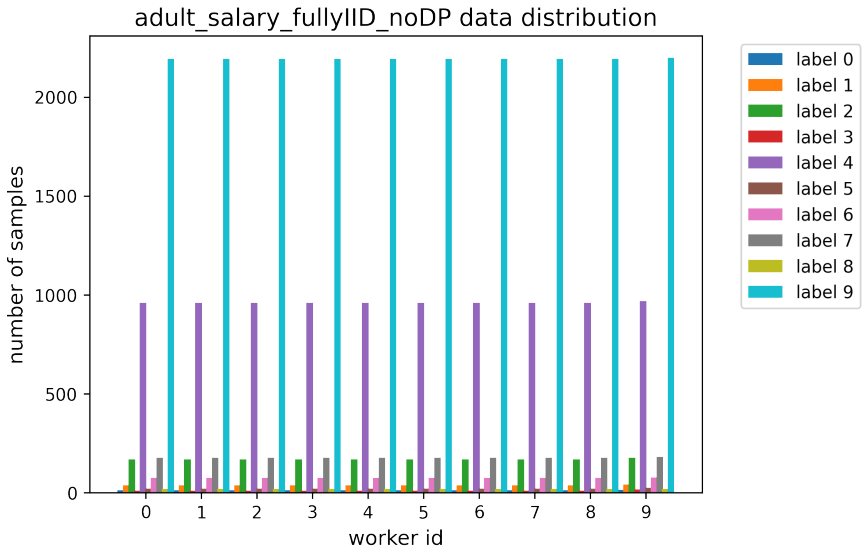

Figure 4.2: Number of samples per label on 10 workers nodes in fully IID data distribution scenario

4.3.2 Scenario 2: 30% 2-class non-IID data distribution

Figure 4.3 gives the number of sampler per label of both IID and non-IID part of the adult dataset. The adult dataset is split based on a proportion of 70/30 on IID and nonIID parts at first. In figure 4.3, worker 0 means IID partition and worker 1 means non-IID partition of the adult dataset. Figure 4.4 shows the 30% 2-class nonIID data distribution scenario of adult dataset among 10 worker nodes. In figure 4.4, worker 0-9 means node 1-10 in the whole federated learning setting.

4.3.3 Scenario 3: 50% 2-class nonIID data distribution

Figure 4.5 gives the number of sampler per label of both IID and non-IID part of the adult dataset. The adult dataset is split based on a proportion of 50/50 on IID and nonIID parts at first. In figure 4.5, worker 0 means IID partition and worker 1 means non-IID partition of the adult dataset. Figure 4.6 shows the 50% 2-class IID data distribution scenario of adult dataset among 10 worker nodes. In figure 4.6, worker 0-9 means node 1-10 in the whole federated learning setting.

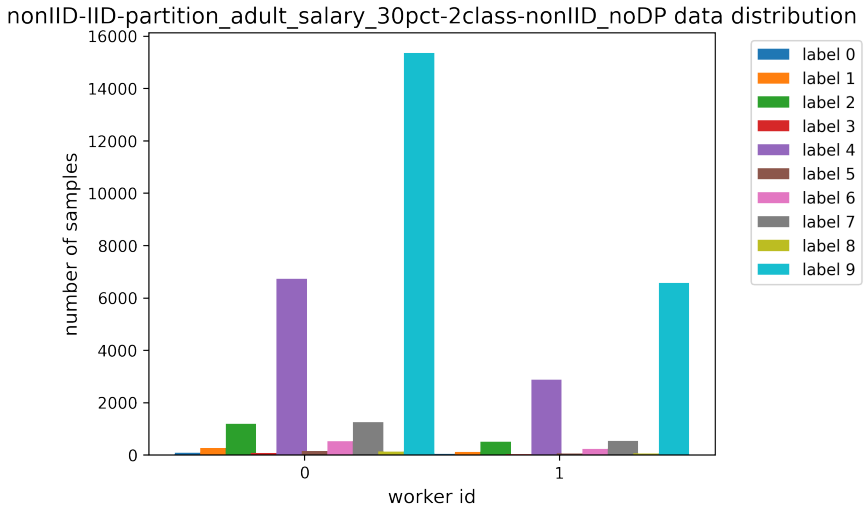

Figure 4.3: Number of samples per label in 30% 2-class nonIID data distribution scenario, node 0 for IID part and node 1 for non-IID part

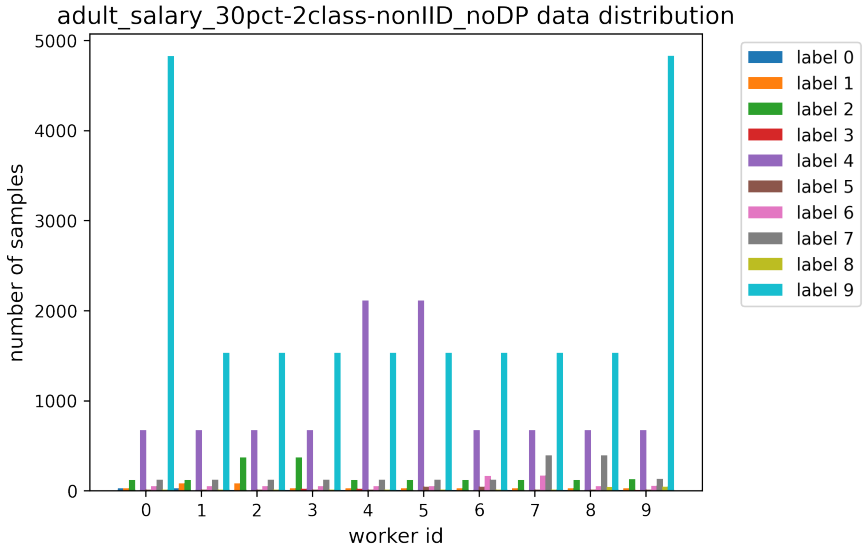

Figure 4.4: Number of samples per label on 10 workers nodes in 30% 2-class nonIID data distribution scenario

4.3.4 Scenario 4: 70% 2-class nonIID data distribution

Figure 4.7 gives the number of sampler per label of both IID and non-IID part of the adult dataset. The adult dataset is split based on a proportion of 30/70 on IID and nonIID parts at first. In figure 4.7, worker 0 means IID partition and worker 1 means

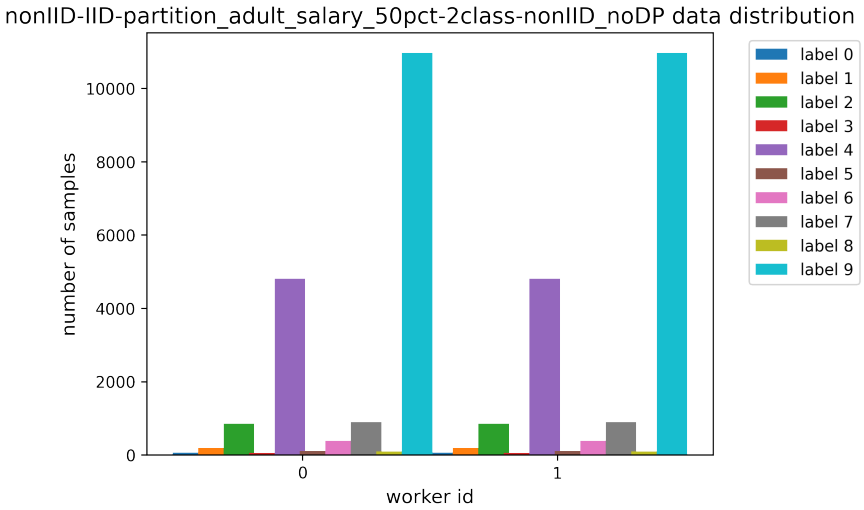

Figure 4.5: Number of samples per label in 50% 2-class nonIID data distribution scenario, node 0 for IID part and node 1 for non-IID part

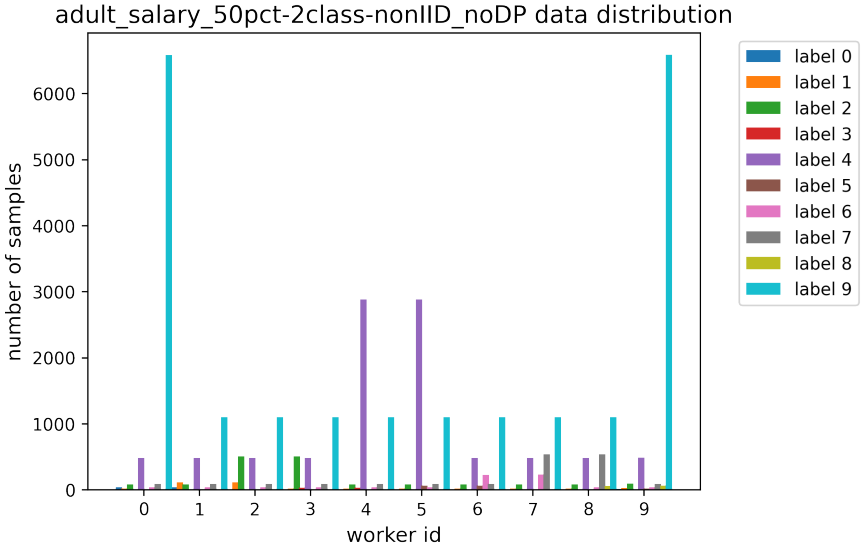

Figure 4.6: Number of samples per label on 10 workers nodes in 50% 2-class nonIID data distribution scenario

non-IID partition of the adult dataset. Figure 4.8 shows the 50% 2-class IID data distribution scenario of adult dataset among 10 worker nodes. In figure 4.8, worker 0-9 means node 1-10 in the whole federated learning setting.

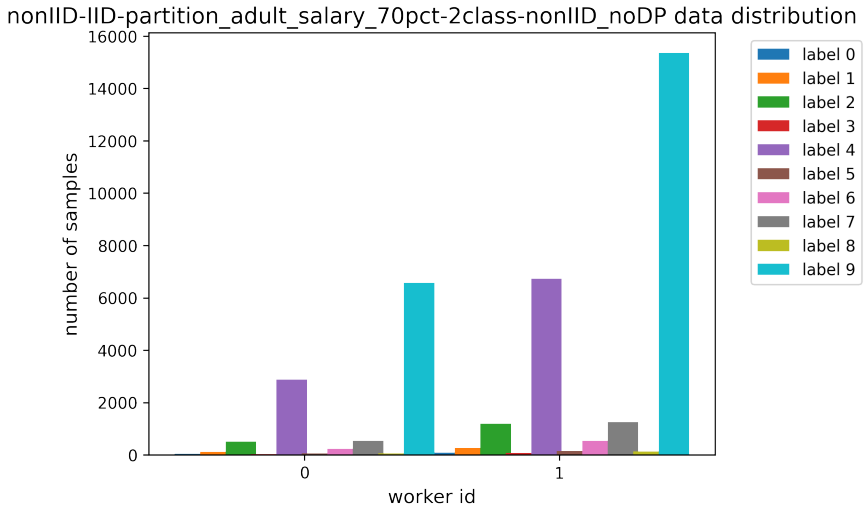

Figure 4.7: Number of samples per label in 70% 2-class nonIID data distribution scenario, node 0 for IID part and node 1 for non-IID part

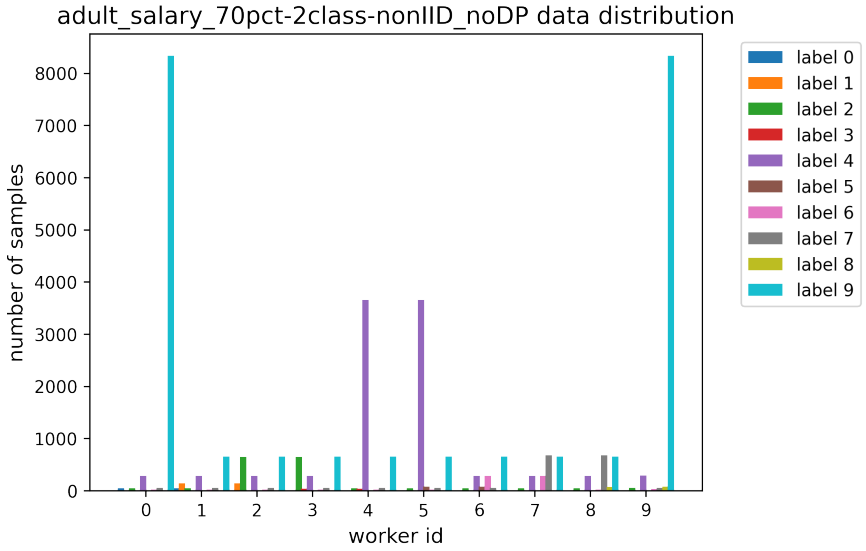

Figure 4.8: Number of samples per label on 10 workers nodes in 70% 2-class nonIID data distribution scenario

4.3.5 Scenario 5: Fully 2-class nonIID data distribution

Figure 4.9 shows the fully 2-class nonIID data distribution scenario of adult dataset among 10 worker nodes. In this extreme scenario, each worker node only has samples from 2 labels. The number of samples in total and per label are different among worker

nodes..

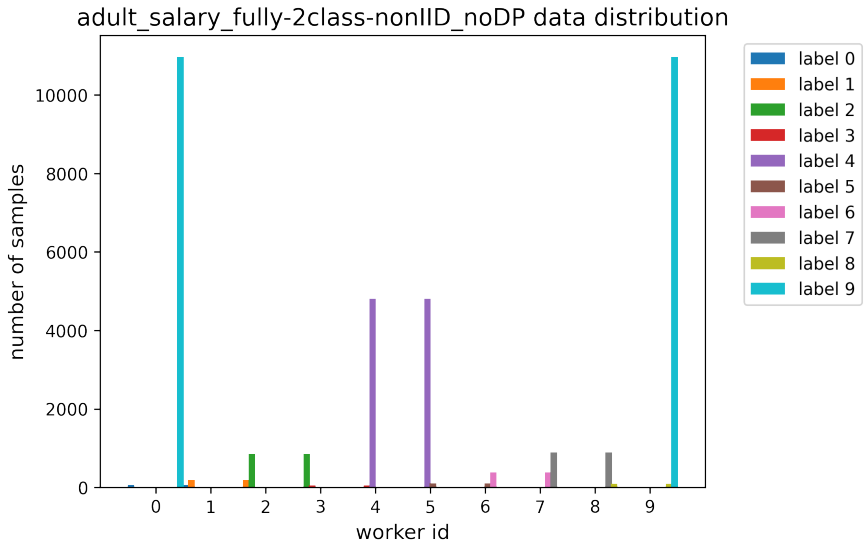

Figure 4.9: Number of samples per label on 10 workers nodes in fully 2-class nonIID data distribution scenarios

4.3.6 Scenario 6: Normal distribution

Table 4.2 shows the μ and σ of each label. The number of samples being allocated on each worker node is decided by the number of data points that fall the specific interval of a generated normal distribution.

Figure 4.10 shows the normal distribution scenario of adult dataset among 10 worker nodes.

4.4 Differential Privacy

Our federated learning framework is developed using PyTorch and PyTorch distributed, so we introduce Opacus to be in charge of the noise adding scheme of differential privacy. Opacus is the PyTorch implementation of DPSGD (13) and Opacus is released by Facebook for high-speed large-scale distributed setting with differential privacy (53). Within Opacus, users need to define the value of target epsilon ϵ and target delta δ on each worker node, and Opacus could compute the noise multiplier and add noise to the raw data. After each training round, Opacus can calculate the privacy spent. The

Sex_Race label	μ	σ
Female_Amer-Indian-Eskimo	0	1
Female_Asian-Pac-Islander	0	0.5
Female_Black	0	0.5
Female_Other	0	0.5
Female_White	0	0.5
Male_Amer-Indian-Eskimo	0	1
Male_Asian-Pac-Islander	0	1
Male_Black	0	1
Male_Other	0	1
Male_White	0	1

Table 4.2: Normal distribution parameters for each label

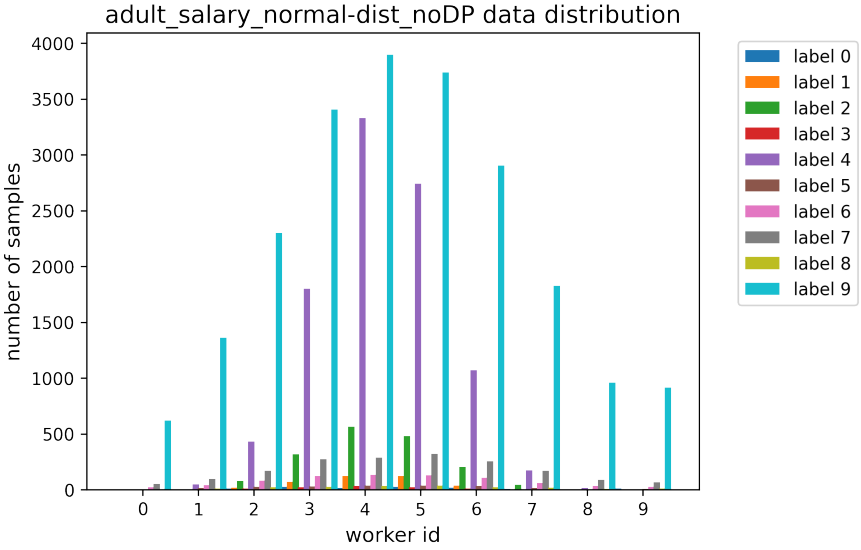

Figure 4.10: Number of samples per label on 10 workers nodes in normal distribution scenarios

higher privacy budget we set, the lower privacy level the dataset has. It means that relatively lower ϵ indicates that there will be more noise added to the raw data, leading to a relatively higher privacy level of the dataset. Detailed integration of Opacus and our federated learning framework can be found in our github repository.

In order to show the impact of privacy in a continuous manner, we performed experiments without Differential Privacy (DP) and conducted experiments with different

budget of low, mid, and high level. The following values of ϵ are chosen for our experiments: 0.1, 0.2, 0.5, 0.8, 1, 1.2, 1.5, 2, 10, 100. The detailed ϵ values within each privacy budget category is shown in Table 4.3.

Privacy Budget Category	ϵ values
Low	0.1, 0.2, 0.5, 0.8
Mid	1, 1.2, 1.5
High	2, 10, 100

Table 4.3: ϵ values in privacy budget categories

4.5 Testing Scheme

4.5.1 Measuring Fairness by Performance of Subgroups

In order to see the performance from both local worker nodes and central server node, the following testing scheme is used in our experiments.

The complete test set of adult dataset contains 9214 samples from 10 labels. Every worker node will do a test on this to record the overall performance of the model per round per node. This is to examine the overall performance of each worker node when they might see some samples which is not included in the feature space of the train set.

Aside from that, the train set distributed on each worker node will also be split as 80/20 for train and validation set. The worker nodes would also do a test on their own validation set to measure the performance when they only see samples with existing feature space.

Also, since we want to measure the performance per label and per class, all of the evaluation metrics will be generated with both overall and per-subgroup values.

4.5.2 Measuring Utility by 4 Metrics

When measuring the performance of a Neural Network classification model based on the Adult dataset, we use the evaluation metrics which are derived from the confusion matrix. Within the confusion matrix of a classification task, True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN) are defined as below:

- True Positive (TP): Number of samples being predicted as positive which are actually positive.
- False Positive (FP): Number of samples being predicted as positive which are actually negative.
- False Negative (FN): Number of samples being predicted as negative which are actually positive.
- True Negative (TN): Number of samples being predicted as negative which are actually negative.

The evaluation metrics we used in the experiments are defined as below:

- Accuracy: The fraction of samples being correctly classified:

$$Accuracy = \frac{TP \ + \ TN}{TP \ + \ TN \ + \ FP \ + \ FN}$$

- Precision: The proportion of the outcomes that are relevant:

$$Precision = \frac{TP}{TP \ + \ FP}$$

- Recall: The proportion of total relevant outcomes correctly predicted:

$$Recall = \frac{TP}{TP \ + \ FN}$$

- F-score (F_1 score): The proportion of the outcomes that are relevant:

$$F_1 score = 2 * \frac{precision \ * \ recall}{precision \ + \ recall}$$

Since this thesis aims at investigating the effect of data bias, we use the weighted averaging scheme to calculate the metrics, specifically F_1 score, precision, and recall. The weighted averaging scheme calculates the metrics for each label and find the average weighted total score based on the support. Here support refers to the number of true samples for each label. The weighted averaging scheme is an improvement from the macro averaging scheme, which takes the label imbalance of the dataset into account. The weighted F_1 score could result in a value which is not in between the weighted precision value and weighted recall value.

5

Experiment Results Analysis

Detailed experiment results about overall, per-class, and per-label performance of all experiments can be found in Appendix A and B. This chapter investigates the effect of various data distribution scenarios and different privacy budget ϵ values in a differentially private federated learning (DPFL) framework based on the vertical comparison and horizontal comparison among all experiment results.

As a reminder of Chapter 4, there are 6 types of data distributions: (1) fully IID, (2) 30% 2-class non-IID, (3) 50% 2-class non-IID, (4) 70% 2-class non-IID, (5) fuly 2-class non-IID, (6) normal distribution. The 10 ϵ values chosen for privacy budget of Differential Privacy (DP) are: 0.1, 0.2, 0.5, 0.8, 1, 1.2, 1.5, 2, 10, 100. We use "worker nodes" and "local nodes" as the same meaning with the definition of clients in federated learning. And the final experiment results are aggregated from 10 collaboratively training worker nodes by *FedAvg* fusion scheme on the central server node.

As mentioned in Table 3.1, the target classes of adult dataset are consisted of *<=50K* and *>50K*. And the Sex_Race labels in adult dataset refer to the combination of Sex and Race features. We regard the *<=50K* target class as the majority class and treat *>50K* target class as the minority class in all sections below. Moreover, the term "under-represented groups" and "minority groups" refer to the minority labels with a small proportion of samples in the whole dataset, specifically *Female_ Amer-Indian-Eskimo* and *Female_ Other*. The term "over representative groups" and "majority groups" refer to the majority labels with a huge percentage of the whole dataset, specifically *Female_ White* and *Male_ White*.

When examining the effect of various data distribution scenarios, we consider the hypothesis that the higher level of data imbalance leads to a better overall performance and worse fairness of under-represented groups. During the investigation on the effect of different privacy budget, we consider the hypothesis that the higher privacy level leads to a better overall performance and worse fairness of under-represented groups.

5.1 Effect of Various Data Distribution Scenarios on Differentially Private Federated Learning

Regarding the differential privacy mechanism in a federated learning setting, we chose one ϵ value from each privacy budget category to investigate the effect of various data distributions on that differentially private scenario. Specifically, we chose the no DP setting, $\epsilon = 0.2$ in low privacy budget, $\epsilon = 1.2$ in mid privacy budget, and $\epsilon = 100$ in high privacy budget. By diving into the overall, per-class, and per-label model performance among various data distribution scenarios within each privacy budget category, we can examine the validity of the hypothesis that a higher level of data bias leads to a better overall performance and worse fairness of under-represented groups.

5.1.1 Baseline Experiment

Since most studies in federated learning assume that the training data is identically independent distributed (IID) among all worker nodes, we use the experiment results of fully IID data distribution with different DP settings as the baseline to investigate the effect of various data distribution scenarios in DPFL.

Data distribution Figure 4.2 shows the number of samples per Sex_Race label under the fully IID data distribution. Since the samples of each label is evenly distributed among 10 worker nodes, this fully IID data distribution scenario has the lowest level of data bias among all 6 data distribution scenarios.

Performance difference between imbalanced target classes Table 5.1 and 5.2 show the final accuracy and F_1 score of the overall dataset and per-class subgroups in the fully IID data distribution scenario with different DP settings. The performance

difference between the *<=50K* target class and *>50K* target class is also included in these two tables.

	No DP	Low ($\epsilon = 0.2$)	Mid ($\epsilon = 1.2$)	High ($\epsilon = 100$)
Overall	0.832429	0.759171	0.827871	0.831452
<=50K	0.909340	0.885520	0.935326	0.928396
>50K	0.599475	0.376476	0.502405	0.537822
Diff (<=50K, >50K)	0.309865	0.509044	0.432921	0.390574

Table 5.1: Overall and per-class final accuracy in baseline experiments, fully IID data distribution scenario with 4 representative DP settings

	No DP	Low ($\epsilon = 0.2$)	Mid ($\epsilon = 1.2$)	High ($\epsilon = 100$)
Overall	0.828505	0.745091	0.816664	0.822951
<=50K	0.890822	0.846828	0.890952	0.892265
>50K	0.639757	0.436945	0.591658	0.613008
Diff (<=50K, >50K)	0.251065	0.409883	0.299294	0.279257

Table 5.2: Overall and per-class final F_1 score in baseline experiments, fully IID data distribution scenario with 4 representative DP settings

As shown in Table 5.1 and 5.2, there is a huge performance difference between target class *<=50K* and *>50K*. This is caused by the noticeable class imbalance within the Adult dataset. Since target class *<=50K* has more than 75% of the samples in the whole dataset, the neural networks model in the binary classification task tends to recognize the target class *<=50K* more, and thus gives less attention to the *>50K*. As a result, the model performance of target class *<=50K* is significantly better than the target class *>50K*. Moreover, since the overall accuracy and F_1 score are both calculate in a weighted averaging scheme, we can see the overall model performance lies in between the performance of these two target classes but obviously inclines more to the model performance of target class *<=50K*.

Performance difference between minority and majority groups Table 5.3 and 5.4 show the final accuracy and F_1 score of the overall dataset and per-label subgroups in the fully IID data distribution scenario with different DP settings.

	No DP	Low ($\epsilon = 0.2$)	Mid ($\epsilon = 1.2$)	High ($\epsilon = 100$)
Overall	0.832429	0.759171	0.827871	0.831452
Female_Amer-Indian-Eskimo	0.794118	0.676471	0.735294	0.852941
Female_Other	0.821429	0.821429	0.892857	0.642857
Female_White	0.822453	0.765073	0.819543	0.820374
Male_White	0.836463	0.754786	0.828624	0.833364

Table 5.3: Overall and per-label final accuracy in baseline experiments, fully IID data distribution scenario with 4 representative DP settings

	No DP	Low ($\epsilon = 0.2$)	Mid ($\epsilon = 1.2$)	High ($\epsilon = 100$)
Overall	0.828505	0.745091	0.816664	0.822951
Female_Amer-Indian-Eskimo	0.798155	0.644752	0.719781	0.847130
Female_Other	0.820252	0.805322	0.878419	0.589286
Female_White	0.817943	0.754992	0.804902	0.811914
Male_White	0.832622	0.739257	0.818001	0.824720

Table 5.4: Overall and per-label final F_1 score in baseline experiments, fully IID data distribution scenario with 4 representative DP settings

As illustrated in Table 5.3 and 5.4, it is clear that the model performance of minority group Sex_Race label *Female_ Amer-Indian-Eskimo* is significantly worse, compared with the overall and other Sex_Race labels in the no DP, low privacy budget and mid privacy budget situations. And the model performance is significantly worse of minority group Sex_Race label *Female_ Other* in the high privacy budget situation. The worse performance on minority groups, compared with overall and majority groups, is caused by the noticeable label imbalance within the Adult dataset. Sex_Race label *Female_ Amer-Indian-Eskimo* and *Female_ Other* only have 0.36% and 0.29% of the samples in the whole dataset. And their proportion of target class *<50K* are 8.43% and 6.67%, which are both significantly lower than the overall *<50K* proportion of 24.8%. As a result, the model would naturally give worse performance on minority groups. On the other side, the model performance of majority groups in Sex_Race label does not show substantial difference compared with the overall model performance.

5.1.2 Extreme Cases

Here we consider the fully 2-class non-IID and normal data distribution with different DP settings as extreme cases when investigating the effect of various data distribution scenarios on DPFL.

Data Distribution Figure 4.9 and 4.10 show the number of samples per Sex_Race label among all worker nodes in fully 2-class non-IID and normal data distribution scenarios. The fully 2-class non-IID data distribution scenario only allocates samples of two Sex_Race labels to each worker node. And the normal data distribution scenario allocates most of the samples in one Sex_Race in a limited number of worker nodes. As a result, some Sex_Race labels are even invisible for particular worker nodes, leading to a substantial high level of data bias.

Performance difference between imbalanced target classes Table 5.5 and 5.6 show the overall and per-class performance difference between extreme cases and baseline experiments. To be more specific, these two tables present the difference of final accuracy and F_1 score on the overall dataset and per-class subgroups by calculating the difference value between extreme cases (fully 2-class non-IID and normal data distribution) and baseline experiments.

	No DP	Low ($\epsilon = 0.2$)	Mid ($\epsilon = 1.2$)	High ($\epsilon = 100$)
Diff_overall (fully non-IID, baseline)	+0.005861	+0.024745	+0.000868	-0.001953
Diff_overall (normal, baseline)	+0.002062	+0.023768	+0.001085	-0.005209
Diff_<=50K (fully non-IID, baseline)	+0.014725	-0.023819	-0.009384	-0.000722
Diff_<=50K (normal, baseline)	+0.010827	-0.044030	-0.012271	-0.003031
Diff_>50K (fully non-IID, baseline)	-0.020988	+0.171841	+0.031919	-0.005684
Diff_>50K (normal, baseline)	-0.024486	+0.229121	+0.041539	-0.011805

Table 5.5: Overall and per-class final accuracy in extreme cases compared with baseline experiments, fully 2-class non-IID and normal data distribution scenario with 4 representative DP settings

Looking at Table 5.5 and 5.6, it is apparent that the difference of final accuracy on target class *>50K* is significantly larger and the overall difference is substantially lower in all DP settings. And the difference of final F_1 score shows a significantly larger difference on target class *>50K* with a substantially lower difference on target class

	No DP	Low ($\epsilon = 0.2$)	Mid ($\epsilon = 1.2$)	High ($\epsilon = 100$)
Diff_overall (fully non-IID, baseline)	+0.003700	+0.037605	+0.003604	-0.002201
Diff_overall (normal, baseline)	+0.000070	+0.040750	+0.004596	-0.005586
Diff_<=50K (fully non-IID, baseline)	+0.004924	+0.010233	-0.000490	-0.001188
Diff_<=50K (normal, baseline)	+0.002333	+0.006738	-0.000671	-0.003284
Diff_>50K (fully non-IID, baseline)	-0.000008	+0.120512	+0.016000	-0.005268
Diff_>50K (normal, baseline)	-0.006785	+0.143768	+0.020547	-0.012559

Table 5.6: Overall and per-class final F_1 score in extreme cases compared with baseline experiments, fully 2-class non-IID and normal data distribution scenario with 4 representative DP settings

<=50K in all privacy budget categories. These results indicate that the change in data bias level has larger impact on the model performance of target class *>50K*.

Performance difference between minority and majority groups Table 5.7 and 5.8 show the overall and per-label performance difference between extreme cases and baseline experiments. To be more specific, these two tables present the difference of final accuracy and F_1 score on the overall dataset and per-class subgroups by calculating the value difference between these two items: (1) performance difference in an extreme data distribution (fully 2-class non-IID or normal data distribution) calculated by the per-label performance value minus overall performance value, and (2) performance difference in the baseline experiment (fully IID) calculated by the per-label performance value minus overall performance value. Besides, we use abbreviations to represent Sex_Race labels in these two tables, specifically FA for *Female_ Amer-Indian-Eskimo*, FO for *Female_ Other*, FW for *Female_ White*, and MW for *Male_ White*.

Since we compare the FA accuracy and overall accuracy between extreme cases and the baseline experiments in Table 5.7 and 5.8, it is clear that the difference between extreme cases and baseline regarding FA-overall and FO-overall accuracy and F_1 score is significantly larger, compared with the FW-overall and MW-overall performance difference. Thus, we can see that switch of distribution scheme (from baseline to extreme cases) has larger impact on the performance of minority groups (Sex_Race label *Female_ Amer-Latin-Eskimo* and *Female_ Other*). On the other hand, the majority groups (Sex_Race label *Female_ White* and *Male_ White*) have shown no big change when the data distribution scenario alters.

	No DP	Low ($\epsilon = 0.2$)	Mid ($\epsilon = 1.2$)	High ($\epsilon = 100$)
Diff_FA-overall (fully non-IID, baseline)	+0.111786	+0.181137	+0.146191	-0.115694
Diff_FA-overall (normal, baseline)	+0.027349	+0.093879	-0.030497	-0.024203
Diff_FO-overall (fully non-IID, baseline)	-0.041576	-0.024745	-0.108011	+0.251953
Diff_FO-overall (normal, baseline)	+0.140795	-0.095197	+0.034629	+0.148066
Diff_FW-overall (fully non-IID, baseline)	+0.018671	-0.006450	-0.005858	+0.015675
Diff_FW-overall (normal, baseline)	+0.015402	-0.007552	+0.008478	+0.009367
Diff_MW-overall (fully non-IID, baseline)	-0.008413	+0.004425	+0.002231	-0.004064
Diff_MW-overall (normal, baseline)	-0.004979	+0.005038	-0.000356	-0.002266

Table 5.7: Final accuracy difference between overall dataset and per-label subgroups, comparing extreme cases with baseline experiments, fully 2-class non-IID and normal data distribution scenario with 4 representative DP settings

	No DP	Low ($\epsilon = 0.2$)	Mid ($\epsilon = 1.2$)	High ($\epsilon = 100$)
Diff_FA-overall (fully non-IID, baseline)	+0.108488	+0.206221	+0.154785	-0.141266
Diff_FA-overall (normal, baseline)	+0.010033	+0.112653	-0.035652	-0.022200
Diff_FO-overall (fully non-IID, baseline)	-0.082881	-0.039897	-0.122283	+0.296108
Diff_FO-overall (normal, baseline)	+0.142737	-0.102780	+0.043969	+0.202014
Diff_FW-overall (fully non-IID, baseline)	+0.019589	-0.011142	-0.004212	+0.015669
Diff_FW-overall (normal, baseline)	+0.017117	-0.010286	+0.010947	+0.010203
Diff_MW-overall (fully non-IID, baseline)	-0.009254	+0.006319	+0.002101	-0.003906
Diff_MW-overall (normal, baseline)	-0.005098	+0.005669	-0.000872	-0.002988

Table 5.8: Final F_1 score difference between overall dataset and per-label subgroups, comparing extreme cases with baseline experiments, fully 2-class non-IID and normal data distribution scenario with 4 representative DP settings

5.1.3 Middle Cases

After diving into extreme cases of fully 2-class non-IID and normal data distribution, we also want to analyze the performance difference along with a continuous changing of the data distribution scenario. Here we use the partial non-IID data distribution scheme, and compare the overall, per-class, and per-label model performance of 0%, 30%, 50%, 70%, and 100% 2-class non-IID data distribution scenario with different DP settings.

Data Distribution Figure 4.2, 4.4, 4.6, 4.8, and 4.9 show the number of samples per Sex_Race label under different data distribution scenarios with different non-IID proportions. As we can see from the figures, an increasing non-IID proportions of

the whole training set from 0% to 100% leads to an increasing level of data bias in experiments.

Performance difference between imbalanced target classes Figure 5.1 and 5.2 present the change of overall and per-class model performance without DP along with an increasing non-IID proportion in the training set. Likewise, Figure 5.3 and 5.4 present the performance change with privacy budget $\epsilon = 0.2$, Figure 5.5 and 5.6 present the performance change with privacy budget $\epsilon = 1.2$, and Figure 5.7 and 5.8 present the change with privacy budget $\epsilon = 100$ along with an increasing non-IID proportion in the training set. Aside from the trend shown in figures above, Table 5.9 and 5.10 show the overall and per-class standard deviation and range in the change of non-IID proportion in the training set.

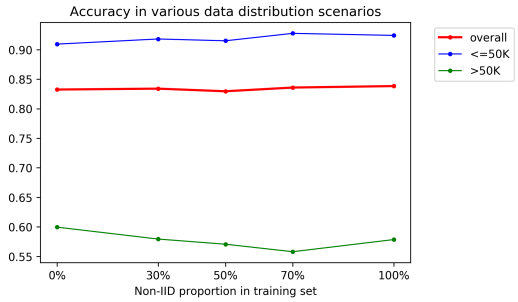

Figure 5.1: Overall and per-class accuracy in different non-IID proportion of the training set, without DP

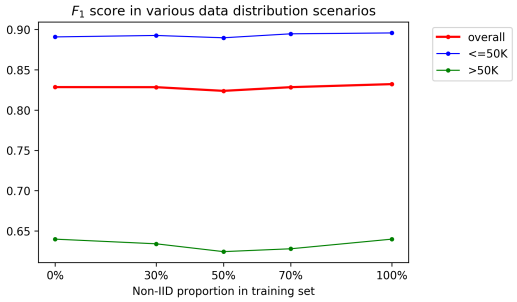

Figure 5.2: Overall and per-class F_1 score in different non-IID proportion of the training set, without DP

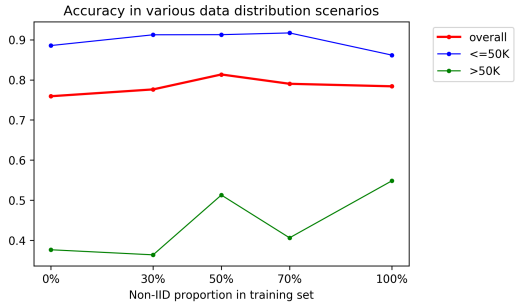

Figure 5.3: Overall and per-class accuracy in different non-IID proportion of the training set, privacy budget $\epsilon = 0.2$

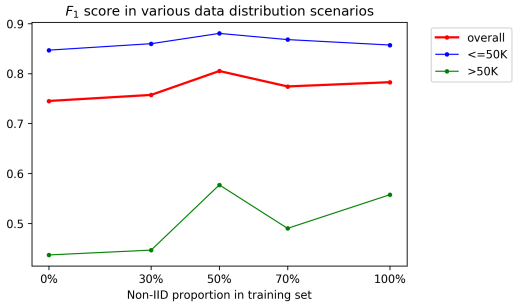

Figure 5.4: Overall and per-class F_1 score in different non-IID proportion of the training set, privacy budget $\epsilon = 0.2$

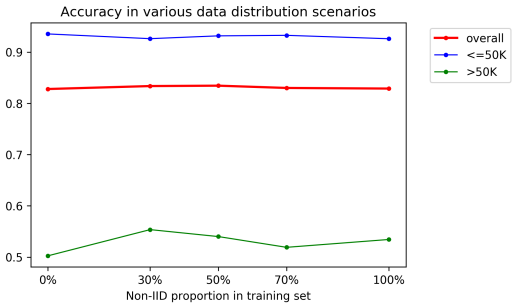

Figure 5.5: Overall and per-class accuracy in different non-IID proportion of the training set, privacy budget $\epsilon = 1.2$

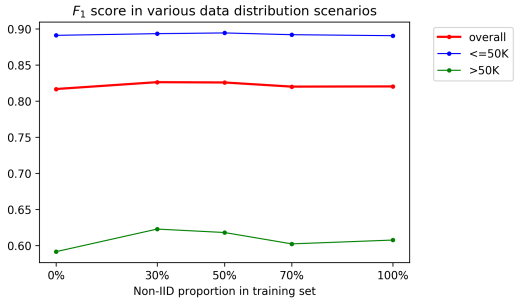

Figure 5.6: Overall and per-class F_1 score in different non-IID proportion of the training set, privacy budget $\epsilon = 1.2$

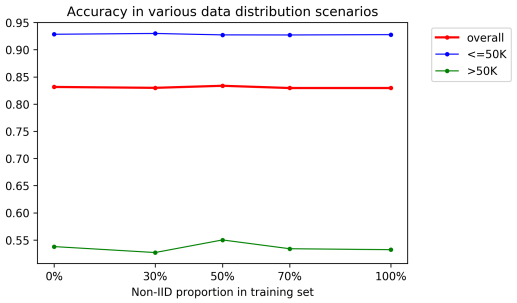

Figure 5.7: Overall and per-class accuracy in different non-IID proportion of the training set, privacy budget $\epsilon = 100$

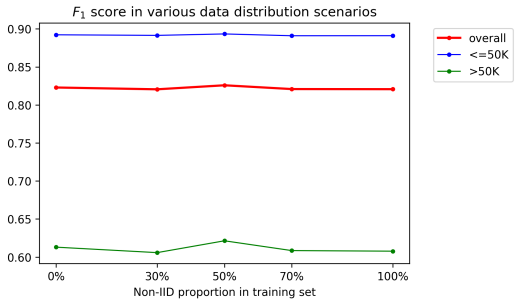

Figure 5.8: Overall and per-class F_1 score in different non-IID proportion of the training set, privacy budget $\epsilon = 100$

	No DP	Low ($\epsilon = 0.2$)	Mid ($\epsilon = 1.2$)	High ($\epsilon = 100$)
StdDev_overall	0.002725	0.016242	0.002500	0.002232
Range_overall	0.008791	0.054265	0.006511	0.007380
StdDev_<=50K	0.005921	0.028584	0.004332	0.001356
Range_<=50K	0.018190	0.075501	0.012271	0.004475
StdDev_>50K	0.012389	0.091695	0.016937	0.008053
Range_>50K	0.041539	0.241802	0.051159	0.024049

Table 5.9: Standard deviation and range of overall and per-label final accuracy in the change of non-IID proportion in training set, with 4 representative DP settings

As we can see from figures above, the overall and per-class model performance change are all not monotonic along with the increasing non-IID proportion of the training set. Besides, the results shown in these figures directly tell us that the huge performance

	No DP	Low ($\epsilon = 0.2$)	Mid ($\epsilon = 1.2$)	High ($\epsilon = 100$)
StdDev_overall	0.002426	0.019547	0.003330	0.002583
Range_overall	0.008368	0.059971	0.009490	0.008509
StdDev_<=50K	0.002067	0.010758	0.001473	0.001341
Range_<=50K	0.006015	0.033489	0.003982	0.004405
StdDev_>50K	0.005711	0.059682	0.010265	0.006494
Range_>50K	0.015504	0.143768	0.031220	0.020939

Table 5.10: Standard deviation and range of overall and per-label final F_1 score in the change of non-IID proportion in training set, with 4 representative DP settings

difference between target class *<=50K* and *>50K* always exists no matter how the data distribution scenario changes with different non-IID proportions in the training set.

As stated in Table 5.9 and 5.10, the change of non-IID proportion in the training set almost brings no change in the overall model performance in all DP settings except the low privacy budget category ($\epsilon = 0.2$). The DP theory clearly explains this phenomenon. Having an extremely low ϵ value as the privacy budget in training process leads to a substantially high level of privacy. As a result, a huge amount of noise is added to the local raw dataset, making the training set much more indistinguishable to the neural networks model of this binary classification task. Thus, when the non-IID proportion changes in this situation, there is a larger reflection on the data samples allocated to each worker node. Thus, the model performance would significantly change in this high privacy level compared with other privacy budgets categories. Also, a substantially huge increase in the overall model performance with $\epsilon = 0.2$ and 50% non-IID proportion of the training set can be seen from Figure 5.3 and 5.4. Similarly, $\epsilon = 1.2$ shows a sudden increase in 30% non-IID in Figure 5.5 and 5.6, and $\epsilon = 1.2$ presents a sudden increase in 50% non-IID in Figure 5.7 and 5.8.

Aside from the overall model performance, Table 5.9 and 5.10 also indicate that the standard deviation and range of target class *>50K* is significantly higher than the overall performance and the target class *<=50K*. This larger impact of data distribution scenarios changes on target class *>50K* can be explained by the high level of class imbalance within the Adult dataset.

5.1.4 Performance difference between minority and majority groups

Figure 5.9 and 5.10 present the change of overall and per-label model performance without DP along with an increasing non-IID proportion in the training set. Likewise, Figure 5.11 and 5.12 present the performance change with privacy budget $\epsilon = 0.2$, Figure 5.13 and 5.14 present the performance change with privacy budget $\epsilon = 1.2$, and Figure 5.15 and 5.16 present the change with privacy budget $\epsilon = 100$ along with an increasing non-IID proportion in the training set. Aside from the trend shown in figures above, Table 5.11 and 5.12 show the overall and per-class standard deviation and range in the change of non-IID proportion in the training set.

Besides, we use abbreviations to represent Sex_Race labels in these two tables, specifically FA for *Female_ Amer-Indian-Eskimo*, FO for *Female_ Other*, FW for *Female_ White*, and MW for *Male_ White*.

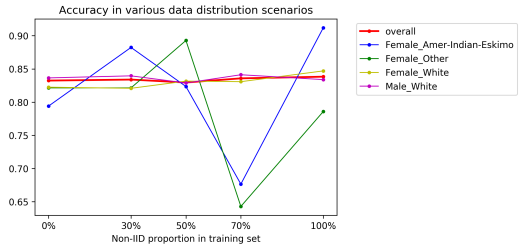

Figure 5.9: Overall and per-label accuracy in different non-IID proportion of the training set, without DP

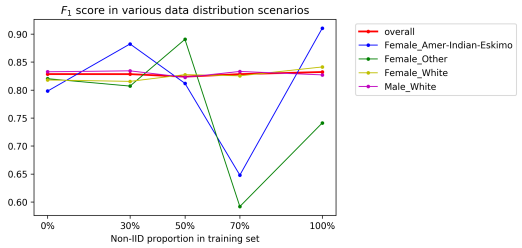

Figure 5.10: Overall and per-label F_1 score in different non-IID proportion of the training set, without DP

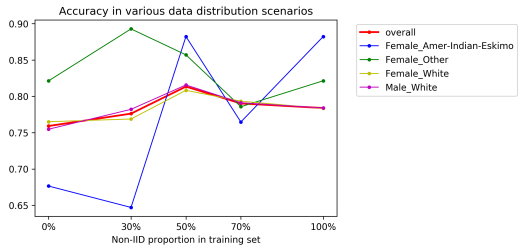

Figure 5.11: Overall and per-label accuracy in different non-IID proportion of the training set, privacy budget $\epsilon = 0.2$

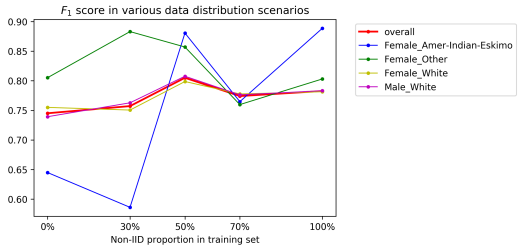

Figure 5.12: Overall and per-label F_1 score in different non-IID proportion of the training set, privacy budget $\epsilon = 0.2$

As shown in figures above, we can see that the model performance of overall dataset and majority groups (Sex_Race label *Female_ White* and *Male_ White*) are quite sim-

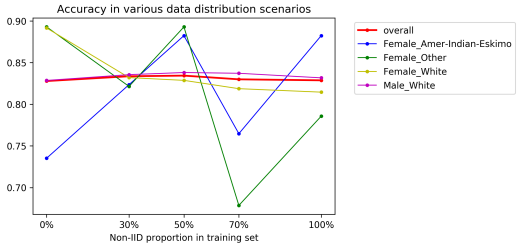

Figure 5.13: Overall and per-label accuracy in different non-IID proportion of the training set, privacy budget $\epsilon = 1.2$

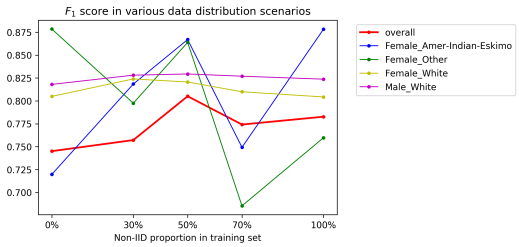

Figure 5.14: Overall and per-label F_1 score in different non-IID proportion of the training set, privacy budget $\epsilon = 1.2$

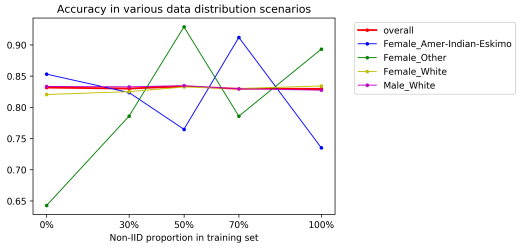

Figure 5.15: Overall and per-label accuracy in different non-IID proportion of the training set, privacy budget $\epsilon = 100$

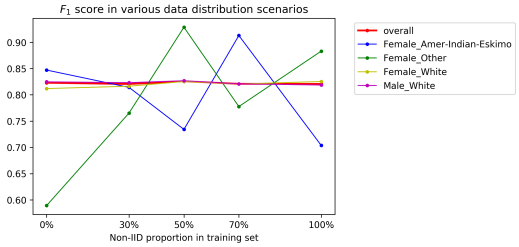

Figure 5.16: Overall and per-label F_1 score in different non-IID proportion of the training set, privacy budget $\epsilon = 100$

	No DP	Low ($\epsilon = 0.2$)	Mid ($\epsilon = 1.2$)	High ($\epsilon = 100$)
StdDev_FA	0.074825	0.090918	0.068802	0.057376
Range_FA	0.235294	0.235294	0.176471	0.176471
StdDev_FO	0.098888	0.046107	0.084179	0.091636
Range_FO	0.321429	0.142857	0.250000	0.285714
StdDev_FW	0.009180	0.014565	0.006443	0.004823
Range_FW	0.026195	0.043243	0.017464	0.013722
StdDev_MW	0.004104	0.017814	0.003722	0.003215
Range_MW	0.012397	0.060893	0.009480	0.008751

Table 5.11: Standard deviation and range of overall and per-label final accuracy in the change of non-IID proportion in training set, with 4 representative DP settings

ilar, demonstrating by the very close standard deviation and range value in the change of non-IID proportions with all DP setting except $\epsilon = 1.2$.

On the other hand, it is clear that the standard deviation of model performance on minority groups (Sex_Race label *Female_Amer-Indian-Eskimo* and *Female_Other*)

	No DP	Low ($\epsilon = 0.2$)	Mid ($\epsilon = 1.2$)	High ($\epsilon = 100$)
StdDev_FA	0.083373	0.112654	0.072210	0.069522
Range_FA	0.262379	0.302605	0.189445	0.209264
StdDev_FO	0.116924	0.049354	0.080661	0.107188
Range_FO	0.371222	0.139901	0.241618	0.339285
StdDev_FW	0.009000	0.016952	0.007931	0.004935
Range_FW	0.025742	0.048281	0.019653	0.013468
StdDev_MW	0.004129	0.021113	0.003933	0.003631
Range_MW	0.011597	0.068427	0.011373	0.010714

Table 5.12: Standard deviation and range of overall and per-label final F_1 score in the change of non-IID proportion in training set, with 4 representative DP settings

are substantially higher than the overall performance and minority groups. Thus, we ca see that the change of non-IID proportion has a larger impact on the minority groups rather than majority groups.

5.1.5 Summary

The validity of the hypothesis that a higher level of data bias leads to a better overall performance and worse fairness of under-represented groups has been proved based on the experiment results in this section.

Since we use the fully IID data distribution scenario as the baseline, we can see that the other 5 data distribution scenarios (30% non-IID, 50% non-IID, 70% non-IID, fully non-IID, and normal data distribution) all get a better overall performance than the baseline in all privacy budget categories (low, mid, high). Therefore, we can conclude that the overall model performance is better when increasing the level of data imbalance.

Also, in the series of partial non-IID experiments, the level of data bias becomes higher when we increase the non-IID proportion in the training set. Along with the increasing non-IID proportion, we can see a significant impact on the model performance of minority groups (Sex_Race label *Female_ Amer-Indian-Eskimo* and *Female_ Other*). At the same time, the model performance of majority groups (Sex_Race label *Female_ White* and *Male_ White*) is not significantly affected by the changes of non-IID proportion. Thus, we can conclude that the model performance of minority groups is largely affected by the variety of data distribution scenarios. In other words, the fairness

of under-represented groups in the whole dataset is affected by various data distribution scenarios.

5.2 Effect of Different Privacy Budget on Differentially Private Federated Learning

Here we would like to quickly recap the differential privacy (DP) mechanism. A lower privacy budget ϵ value results in a higher privacy level. And a higher privacy level leads to more noise added to the raw dataset. Thus, the low, mid, and high privacy budget categories represent the high, mid, and low privacy level, leading to large, modest, and small noise being added to the raw dataset.

Regarding the differential privacy setting in the federated learning setting, we chose 9 several ϵ values from each privacy budget category, specifically 0.2, 0.5, and 0.8 in low privacy budget, 1, 1.2, 1.5 in mid privacy budget, and 2, 10, and 100 in high privacy budget. By diving into the overall, per-class, and per-label model performance difference among different privacy budget ϵ values within each type of data distribution scenarios, we can examine the validity of the hypothesis that a higher privacy level leads to a worse overall performance and worse fairness of under-represented groups.

5.2.1 Performance difference between imbalanced target classes

5.2.1.1 Baseline Experiment

In order to measure the effect of different privacy budgets (ϵ values), we hereby use the experiment results without DP as the baseline. Within the 6 data distribution scenarios in the experiment design, we choose to use the following three representative data distribution scenarios to investigate the effect of different privacy budget (ϵ values): (1) fully IID, (2) fully 2-class non-IID, and (3) normal data distribution.

Table 5.13 and 5.14 show the final accuracy and F_1 score of the overall dataset and per-class subgroups in the 3 representative data distribution scenario without DP. The performance difference between the $<=50K$ target class and $>50K$ target class is also included in these two tables.

As shown in Table 5.13 and 5.14, it is clear that the huge performance difference between target class $<=50K$ and $>50K$ always exists in the baseline experiments.

	Fully IID	Fully non-IID	Normal
Overall	0.832429	0.838290	0.834491
<=50K	0.909340	0.924065	0.920167
>50K	0.599475	0.578487	0.574989
Diff (<=50K, >50K)	0.309865	0.345578	0.345178

Table 5.13: Overall and per-class final accuracy in 3 representative data distribution scenarios (fully IID, fully 2-class non-IID, and normal data distribution) without DP

	Fully IID	Fully non-IID	Normal
Overall	0.828505	0.832205	0.828575
<=50K	0.890822	0.895746	0.893155
>50K	0.639757	0.639749	0.632972
Diff (<=50K, >50K)	0.251065	0.255997	0.260183

Table 5.14: Overall and per-class final F_1 score in baseline experiments, fully IID, fully 2-class non-IID, and normal data distribution scenarios without DP

5.2.1.2 Privacy budget (ϵ value) changes from 0.2 to 100

As we mentioned in Chapter 4, we choose 3 ϵ values for each privacy budget category, specifically 0.2, 0.5, 0.8 for low privacy budget, 1, 1.2, 1.5 for mid privacy budget, and 2, 10, 100 for high privacy budget.

Figure 5.17 and 5.18 present the change of overall and per-class model performance in fully IID data distribution along with an increasing ϵ value as the privacy budget. Likewise, Figure 5.19 and 5.20 illustrate the performance change along with increasing ϵ values in privacy budget within fully 2-class non-IID data distribution. And Figure 5.21 and 5.22 show the performance change along with increasing ϵ values in privacy budget within normal distribution.

Table 5.15 and 5.16 show the final accuracy and F_1 score of the overall dataset and per-class subgroups in 3 representative data distribution scenarios with increasing privacy budget (ϵ changes from 0.2 to 100).

As we can see from Table 5.15 and 5.16, it is clear that the standard deviation on performance of target class *>50K* is significantly higher in fully IID data distribution. And standard deviation on performance of target class *<=50K* is significantly higher in fully 2-class non-IID data distribution. Given that, we can see that there is no general

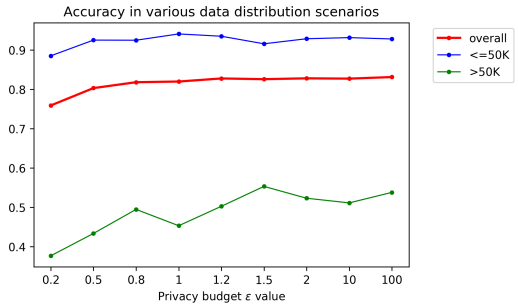

Figure 5.17: Overall and per-label accuracy in 9 different privacy budget ϵ values, within fully IID data distribution scenario

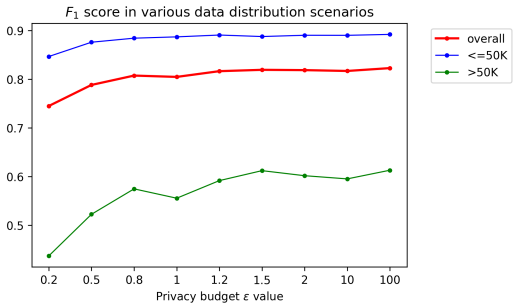

Figure 5.18: Overall and per-label F_1 score in 9 different privacy budget ϵ values, within fully IID data distribution scenario

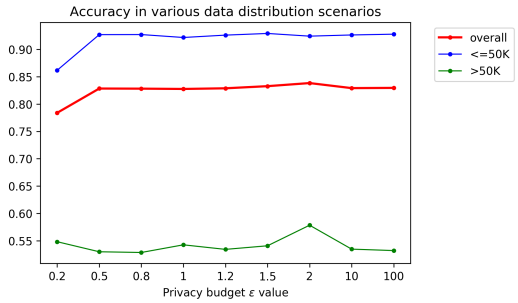

Figure 5.19: Overall and per-label accuracy in 9 different privacy budget ϵ values, within fully 2-class non-IID data distribution scenario

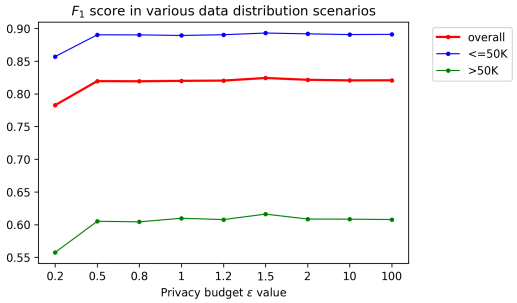

Figure 5.20: Overall and per-label F_1 score in 9 different privacy budget ϵ values, within fully 2-class non-IID data distribution scenario

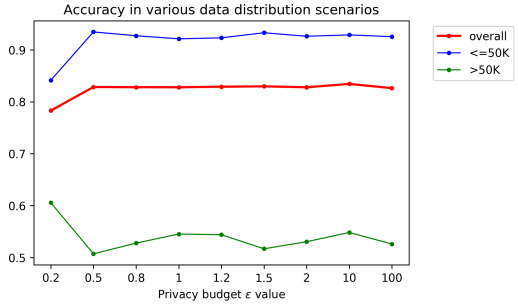

Figure 5.21: Overall and per-label accuracy in 9 different privacy budget ϵ values, within normal data distribution scenario

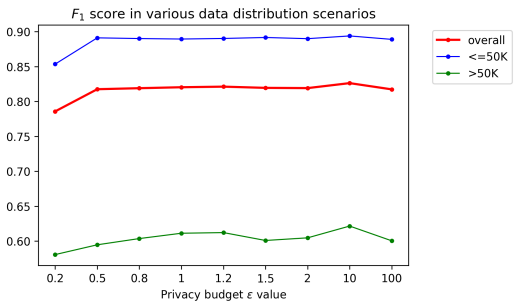

Figure 5.22: Overall and per-label F_1 score in 9 different privacy budget ϵ values, within normal data distribution scenario

	Fully IID	Fully non-IID	Normal
Std_overall	0.021527	0.014333	0.014596
Range_overall	0.072281	0.048730	0.051335
Std_<=50K	0.015193	0.020536	0.027280
Range_<=50K	0.055724	0.067706	0.092969
Std_>50K	0.053103	0.006290	0.026783
Range_>50K	0.176650	0.019677	0.098382

Table 5.15: Standard deviation of overall and per-class final accuracy in 3 representative data distribution scenarios (fully IID, fully 2-class non-IID, and normal data distribution) with ϵ value changes from 0.2 to 100

	Fully IID	Fully non-IID	Normal
Std_overall	0.023271	0.012041	0.011047
Range_overall	0.077860	0.041563	0.040467
Std_<=50K	0.013570	0.010668	0.011751
Range_<=50K	0.045437	0.035945	0.040341
Std_>50K	0.053605	0.016319	0.010934
Range_>50K	0.176063	0.058579	0.040848

Table 5.16: Standard deviation of overall and per-class final F_1 score in 3 representative data distribution scenarios (fully IID, fully 2-class non-IID, and normal data distribution) with ϵ value changes from 0.2 to 100

pattern on the change of overall and per-class performance among different kinds of data distributions. However, if we take a deeper look at figure above, it is clear that when ϵ value changes from 0.2 to 0.5, the performance of overall dataset and target class *<=50K* both has a sudden increase in all of the 3 representative data distribution scenarios. This phenomenon can be explained by the DP theory. When ϵ is 0.2, we have add a huge amount of noise to the dataset, resulting in a highly indistinguishable dataset to local models. Then when we release the privacy limitation to $\epsilon = 0.5$, the local model has a more distinguishable dataset. As a result, the model could have better performance on the overall dataset and the target class *<=50K* with more than 75% of the whole dataset.

Moreover, Figure 5.17 and 5.18 clearly indicate that the change of privacy budget (ϵ value) has a substantially larger impact on the target class *>50K* with less than 25% of

the dataset in fully IID data distribution, compared with the overall dataset and target class *<=50K*.

5.2.2 Performance difference between minority and majority groups

5.2.2.1 Baseline Experiment

In order to measure the effect of different privacy budgets (ϵ values), we hereby use the experiment results without DP as the baseline. 3 out of 6 representative data distribution scenarios are chosen to investigate the effect of different privacy budget (ϵ value): (1) fully IID, (2) fully 2-class non-IID, and (3) normal data distribution.

Table 5.17 and 5.18 show the final accuracy and F_1 score of the overall dataset and per-class subgroups in the 3 representative data distribution scenario without DP.

	Fully IID	Fully non-IID	Normal
Overall	0.832429	0.838290	0.834491
Female_Amer-Indian-Eskimo	0.794118	0.911765	0.823529
Female_Other	0.821429	0.785714	0.964286
Female_White	0.822453	0.846985	0.839917
Male_White	0.836463	0.833911	0.833546

Table 5.17: Overall and per-label final accuracy in 3 representative data distribution scenarios (fully IID, fully 2-class non-IID, and normal data distribution) without DP

	Fully IID	Fully non-IID	Normal
Overall	0.828505	0.832205	0.828575
Female_Amer-Indian-Eskimo	0.798155	0.910343	0.808258
Female_Other	0.820252	0.741071	0.963059
Female_White	0.817943	0.841232	0.835130
Male_White	0.832622	0.827068	0.827594

Table 5.18: Overall and per-label final F_1 score in 3 representative data distribution scenarios (fully IID, fully 2-class non-IID, and normal data distribution) without DP

As presented in Table 5.17 and 5.18, it is clear that the model performance of Sex_Race label *Female_ Amer-Indian-Eskimo* is significantly worse in the fully IID data distribution and model performance of Sex_Race label *Female_ Other* is substantially worse in fully non-IID data distribution. With these results on hand, we can see that

there is no general pattern on the model performance of minority or majority groups among different data distribution scenarios.

5.2.2.2 Privacy budget (ϵ value) changes from 0.2 to 100

Figure 5.23 and 5.24 present the change of overall and per-label model performance in fully IID data distribution along with an increasing ϵ value from 0.2 to 100 as the privacy budget. Likewise, Figure 5.25 and 5.26 present the performance change in fully 2-class non-IID data distribution with an increasing ϵ value. And Figure 5.27 and 5.28 present the performance change in normal data distribution with an increasing ϵ value. Aside from the trend shown in figures above, Table 5.19 and 5.20 show the standard deviation and range of the overall and per-label model performance in the change of an increasing ϵ value as the privacy budget.

Besides, we use abbreviations to represent Sex_Race labels in these two tables, specifically FA for *Female_ Amer-Indian-Eskimo*, FO for *Female_ Other*, FW for *Female_ White*, and MW for *Male_ White*.

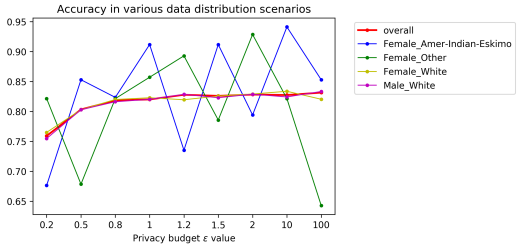

Figure 5.23: Overall and per-label accuracy in fully IID data distribution, with increasing ϵ value from 0.2 to 100

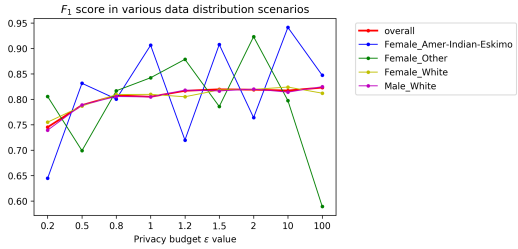

Figure 5.24: Overall and per-label F_1 score in fully IID data distribution, with increasing ϵ value from 0.2 to 100

As shown in figures above, we can see that the model performance of overall dataset and majority groups (Sex_Race label *Female_ White* and *Male_ White*) are quite similar, demonstrating by the very close standard deviation and range value in the change of ϵ value from 0.2 to 100 within all the 3 representative data distribution scenarios.

On the other hand, Table 5.19 and 5.20 clearly shows that the standard deviation of model performance on minority groups (Sex_Race label *Female_ Amer-Indian-Eskimo* and *Female_ Other*) are substantially higher than the overall performance and minority

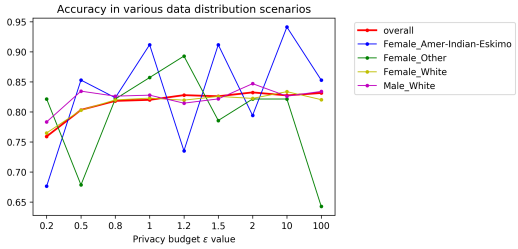

Figure 5.25: Overall and per-label accuracy in fully 2-class non-IID data distribution, with increasing ϵ value from 0.2 to 100

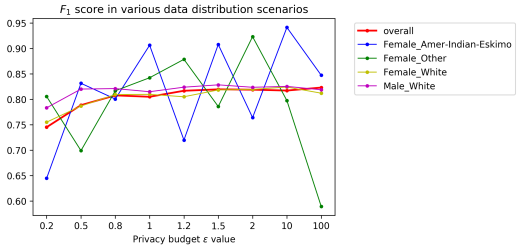

Figure 5.26: Overall and per-label F_1 score in fully 2-class non-IID data distribution, with increasing ϵ value from 0.2 to 100

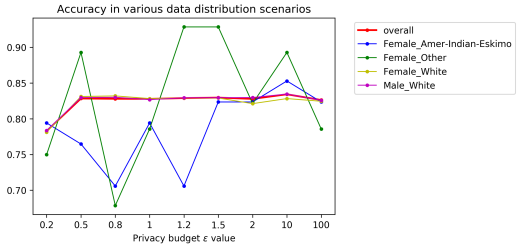

Figure 5.27: Overall and per-label accuracy in normal data distribution, with increasing ϵ value from 0.2 to 100

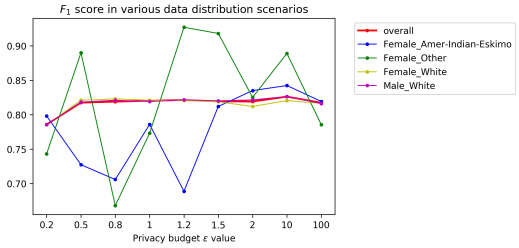

Figure 5.28: Overall and per-label F_1 score in normal data distribution, with increasing ϵ value from 0.2 to 100

	Fully IID	Fully non-IID	Normal
StdDev_FA	0.082025	0.063537	0.049561
Range_FA	0.264705	0.235294	0.147059
StdDev_FO	0.087662	0.075292	0.082096
Range_FO	0.285714	0.250000	0.250000
StdDev_FW	0.019542	0.014718	0.014994
Range_FW	0.068607	0.051143	0.050728
StdDev_MW	0.022760	0.015018	0.014598
Range_MW	0.078578	0.052142	0.051048

Table 5.19: Standard deviation and range of overall and per-label final accuracy in 3 representative data distributions (fully IID, fully 2-class non-IID, and normal data distribution), with ϵ value changes from 0.2 to 100

groups. Thus, we ca see that the change of ϵ value as the privacy budget has a larger impact on the minority groups rather than majority groups and overall dataset.

	Fully IID	Fully non-IID	Normal
StdDev_FA	0.091229	0.070912	0.054218
Range_FA	0.296424	0.266222	0.153672
StdDev_FO	0.092943	0.072208	0.083642
Range_FO	0.333695	0.235877	0.259000
StdDev_FW	0.020179	0.013017	0.011030
Range_FW	0.068979	0.044216	0.037576
StdDev_MW	0.024846	0.012721	0.011289
Range_MW	0.085463	0.045129	0.040857

Table 5.20: Standard deviation and range of overall and per-label final F_1 score in 3 representative data distributions (fully IID, fully 2-class non-IID, and normal data distribution), with ϵ value changes from 0.2 to 100

Moreover, as we can see from the figures, there is no general pattern or monotonic trend in the model performance of minority groups. However, we can see a clear drop on the performance of minority groups when the ϵ decreases from 10 to 100. This phenomenon can be explained by the DP theory. When ϵ has been lifted form 10 to 100, it results in a significantly smaller amount of noise being added to the raw dataset. As a result, less noise in minority groups makes these subgroups with a small number of samples even harder to be correctly classified by the neural network model. Thus, we can see a substantial performance drop on minority groups when ϵ value decreases from 10 to 100.

5.2.3 Summary

The validity of the hypothesis that a higher privacy level leads to a worse overall performance and worse fairness of under-represented groups has been proved based on the experiment results in this section.

Since we use the experiment of fully IID, fully 2-class non-IID, and normal data distribution scenario without DP as the baseline of their own series, we can see that the model performance is worse after introducing DP. Also, a lower privacy budget means a higher privacy level. With more noise being added to the raw dataset (larger ϵ value), the overall model performance is significantly lower. Therefore, we can conclude that the overall model performance is worse when increasing the privacy level.

Also, it is clear that changing privacy budget (ϵ value) has a significant impact on the model performance of minority groups (Sex_Race label *Female_ Amer-Indian-Eskimo* and *Female_ Other*). At the same time, the model performance of majority groups (Sex_Race label *Female_ White* and *Male_ White*) is not significantly affected by the changes of privacy level (ϵ value). Thus, we can conclude that the model performance of minority groups is largely affected by the variety of data distribution scenarios. In other words, the fairness of under-represented groups in the whole dataset is affected by different privacy budgets (ϵ values).

6

Conclusion

In order to investigate the effect of data bias in a differentially private federated learning setting, we designed a comprehensive experiment scheme in this thesis to show the impact of data bias and privacy preserving mechanism on the utility and fairness within the federated learning setting.

Within the experiment scheme, we considered data distribution scenarios, differential privacy budget, federated learning setting, and testing metrics. Specifically, we simulated 6 representative data distributions to mimic real-world machine learning problem situations: (1)fully IID, (2) 30% 2-class non-IID, (3) 50% 2-class non-IID, (4) 70% 2-class non-IID, (5) fully 2-class non-IID, and (6) normal distribution. Aside from the experiments without DP, we chose 3 ϵ values from each privacy budget category to represent different privacy levels required in real-world cases, specifically 0.2, 0.5, and 0.8 in low privacy budget category, 1, 1.2, and 1.5 in mid privacy budget category, and 2, 10, and 100 in high privacy budget category.

In this thesis, 60 experiments are conducted and analyzed, with 6 data distribution scenarios and 10 differential privacy settings. Based on the experiment results, we could draw the following conclusions:

- The general utility, which could be represented by the overall model performance, largely decreases when introducing high privacy level to a differentially private federated learning setting, especially when training on a highly imbalance dataset.

- The fairness of under-represented groups, which could be represented by the per-label model performance, largely decreases when introducing a larger non-IID

proportion in the training set or setting a higher privacy level in a differentially private federated learning setting, especially when training on a highly imbalance dataset.

- The overall model performance is more stable than the per-class and per-label performance in a differentially private federated learning setting.
- When performing a binary classification task on a highly imbalanced dataset, there is a large performance difference between prominent target class and the unapparent target class.
- The model performance of under-represented groups (minority groups) is worse than the majority groups in a highly imbalanced dataset. The larger difference on the number of samples between minority groups and majority groups, the more significant difference will be between the model performance of minority groups and majority groups.

7

Future Work

As future work, there are still several directions worth exploring within the topic of investigating the effect of data bias in differentially private federated learning. The following aspects could be done in the future as an extension of this thesis:

- Implement different federated learning fusion schemes like Krum, Zeno, and Fed+. Design and perform experiments on different FL fusion schemes, so as to introduce the fusion scheme as another controlled variable in the differentially private federated learning setting.
- Design and perform experiments based on other machine learning tasks using neural networks aside from supervised categorical classification tasks. Image classification, next-word prediction, and pattern recognition are all possible tasks.
- Apply the different representative data distribution mechanisms on datasets other than the Adult dataset to see if there are different performance within the dataset when using the data distribution mechanisms on various data formats.

References

[1] European Parliament. General Data Protection Regulation, Regulation (EU) 2016/679. `https://gdpr-info.eu/`, 2018. Online. 1

[2] U.S. Department of Health and Human Services (HHS). Health Insurance Portability and Accountability Act (HIPAA). `https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html`, 1996. Online. 1

[3] 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. 1, 10

[4] 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. *arXiv preprint arXiv:1912.04977*, 2019. 1, 2, 3, 11, 13

[5] Alexandra Moraru, Marko Pesko, Maria Porcius, Carolina Fortuna, and Dunja Mladenic. Using machine learning on sensor data. *Journal of computing and information technology*, 18(4):341–347, 2010. 2

[6] ByungWan Jo and Rana Muhammad Asad Khan. An internet of things system for underground mine air quality pollutant prediction based on azure machine learning. *Sensors*, 18(4):930, 2018. 2

[7] Wesllen Sousa Lima, Hendrio L de Souza Bragança, Kevin G Montero Quispe, and Eduardo J Pereira Souto. Human activity recognition based on symbolic representation algorithms for inertial sensors. *Sensors*, 18(11):4045, 2018. 2

[8] Jae-Neung Lee, Yeong-Hyeon Byeon, Sung-Bum Pan, and Keun-Chang Kwak. An eigenecg network approach based on pcanet for personal identification from ecg signal. *Sensors*, 18(11):4024, 2018. 2

[9] Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*, pages 191–198, 2016. 2

[10] Jidong Chen, Ye Tao, Haoran Wang, and Tao Chen. Big data based fraud risk management at alibaba. *The Journal of Finance and Data Science*, 1(1):1–10, 2015. 2

[11] Yiping Huang, Longmei Zhang, Zhenhua Li, Han Qiu, Tao Sun, Xue Wang, and Helge Berger. Fintech credit risk assessment for smes: Evidence from china. *IMF Working Papers*, 2020(193), 2020. 2

[12] Hamidreza Maharlou, Sharareh R Niakan Kalhori, Shahrbanoo Shahbazi, and Ramin Ravangard. Predicting length of stay in intensive care units after cardiac surgery: comparison of artificial neural networks and adaptive neuro-fuzzy system. *Healthcare informatics research*, 24(2):109–117, 2018. 2

[13] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016. 2, 9, 37

[14] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 486–503. Springer, 2006. 3, 8, 9

[15] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 3, 21

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[17] Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150, 2011. 3

[18] Ali Narin, Ceren Kaya, and Ziynet Pamuk. Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks. *Pattern Analysis and Applications*, pages 1–14, 2021. 3

[19] Avrim Blum, Cynthia Dwork, Frank McSherry, and Kobbi Nissim. Practical privacy: the sulq framework. In *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 128–138, 2005. 7, 8

[20] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006. 8, 9

[21] Cynthia Dwork. A firm foundation for private data analysis. *Communications of the ACM*, 54(1):86–95, 2011. 8

[22] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407, 2014. 8

[23] Federated AI Technology Enabler (FATE) from WeBank's AI team. WeBank and Swiss Re signed Cooperation MoU. `https://www.fedai.org/news/webank-and-swiss-re-signed-cooperation-mou/`, 2019. Website. 11

[24] Swiss Re. Swiss Re Partners with Tencent's WeBank to Research AI Use in Reinsurance. `https://www.swisscham.org/category/bank-finance-insurance/page/2/`, 2019. Website. 12

[25] Pierre Courtiol, Charles Maussion, Matahi Moarii, Elodie Pronier, Samuel Pilcer, Meriem Sefta, Pierre Manceron, Sylvain Toldo, Mikhail Zaslavskiy, Nolwenn Le Stang, et al. Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nature medicine*, 25(10):1519–1525, 2019. 12

[26] Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*, 2018. 12, 15

[27] Timothy Yang, Galen Andrew, Hubert Eichner, Haicheng Sun, Wei Li, Nicholas Kong, Daniel Ramage, and Françoise Beaufays. Applied federated learning: Improving google keyboard query suggestions. *arXiv preprint arXiv:1812.02903*, 2018. 12

[28] Mingqing Chen, Rajiv Mathews, Tom Ouyang, and Françoise Beaufays. Federated learning of out-of-vocabulary words. *arXiv preprint arXiv:1903.10635*, 2019. 12

[29] Swaroop Ramaswamy, Rajiv Mathews, Kanishka Rao, and Françoise Beaufays. Federated learning for emoji prediction in a mobile keyboard. *arXiv preprint arXiv:1906.04329*, 2019. 12, 15

[30] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19, 2019. 13

[31] Fadi Thabtah, Suhel Hammoud, Firuz Kamalov, and Amanda Gonsalves. Data imbalance in classification: Experimental evaluation. *Information Sciences*, 513:429–441, 2020. 14

[32] Sotiris Kotsiantis, Dimitris Kanellopoulos, Panayiotis Pintelas, et al. Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*, 30(1):25–36, 2006. 14

[33] SB Kotsiantis and PE Pintelas. Mixture of expert agents for handling imbalanced data sets. *Annals of Mathematics, Computing & Teleinformatics*, 1(1):46–55, 2003. 14

[34] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002. 14

[35] Pedro Domingos. Metacost: A general method for making classifiers cost-sensitive. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 155–164, 1999. 14

[36] Gary M Weiss. Mining with rarity: a unifying framework. *ACM Sigkdd Explorations Newsletter*, 6(1):7–19, 2004. 14

[37] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009. 14

[38] Robin C Geyer, Tassilo Klein, and Moin Nabi. Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557*, 2017. 15

[39] Jihun Hamm, Yingjun Cao, and Mikhail Belkin. Learning privately from multiparty data. In *International Conference on Machine Learning*, pages 555–563. PMLR, 2016. 15

[40] Moming Duan, Duo Liu, Xianzhang Chen, Yujuan Tan, Jinting Ren, Lei Qiao, and Liang Liang. Astraea: Self-balancing federated learning for improving classification accuracy of mobile deep learning applications. In *2019 IEEE 37th international conference on computer design (ICCD)*, pages 246–254. IEEE, 2019. 15

[41] Lixu Wang, Shichao Xu, Xiao Wang, and Qi Zhu. Addressing class imbalance in federated learning. *arXiv preprint arXiv:2008.06217*, 2020. 15, 16

[42] R Bharat Rao, Sriram Krishnan, and Radu Stefan Niculescu. Data mining for improved cardiac care. *Acm Sigkdd Explorations Newsletter*, 8(1):3–10, 2006. 15

[43] Jiahua Dong, Yang Cong, Gan Sun, and Dongdong Hou. Semantic-transferable weakly-supervised endoscopic lesions segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10712–10721, 2019. 15

[44] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5375–5384, 2016. 15

[45] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 7032–7042, 2017. 15

[46] Tom Farrand, Fatemehsadat Mireshghallah, Sahib Singh, and Andrew Trask. Neither private nor fair: Impact of data imbalance on utility and fairness in differential privacy. In *Proceedings of the 2020 Workshop on Privacy-Preserving Machine Learning in Practice*, pages 15–19, 2020. 17

[47] Felix Sattler, Simon Wiedemann, Klaus-Robert Müller, and Wojciech Samek. Robust and communication-efficient federated learning from non-iid data. *IEEE transactions on neural networks and learning systems*, 31(9):3400–3413, 2019. 17

[48] Qinbin Li, Yiqun Diao, Quan Chen, and Bingsheng He. Federated learning on non-iid data silos: An experimental study. *arXiv preprint arXiv:2102.02079*, 2021. 17

[49] Jonathan Huang. Maximum likelihood estimation of dirichlet distribution parameters. *CMU Technique Report*, 2005. 18

[50] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 21

[51] Ron Kohavi et al. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd*, volume 96, pages 202–207, 1996. 21

[52] Sergey Zagoruyko. PyTorchViz: A small package to create visualizations of PyTorch execution graphs and traces. `https://github.com/szagoruyko/pytorchviz`, 2018. Github. 30

[53] Facebook AI. Introducing Opacus: A high-speed library for training PyTorch models with differential privacy. `https://ai.facebook.com/blog/introducing-opacus-a-high-speed-library-for-training-pytorch-models-with-differential-privacy/`, 2020. Website. 37

Appendices

In order to measure the effect of data bias in differentially private federated learning setting, we conducted experiments with 6 data distribution scenarios and 11 differential privacy settings, by performing a binary classification task using NN on the highly imbalanced Adult dataset.

In order to measure the effect of data bias in differentially private federated learning setting, we conducted experiments with 6 data distribution scenarios and 11 differential privacy settings, by performing a binary classification task using NN on the highly imbalanced Adult dataset.

To be more specific, we performed 66 experiments in total. We use the following 6 data distribution scenarios which have been thoroughly explained in chapter 4: (1) fully IID, (2) 30% 2-class nonIID, (3) 50% 2-class nonIID, (4) 70% 2-class nonIID, (5) fully 2-class nonIID, and (6) normal distribution. And we use a no DP setting and 10 different privacy budget values as listed here: 0.1, 0.2, 0.5, 0.8, 1, 1.2, 1.5, 2, 10, 100.

As described in Chapter 3, we define *<=50K* as the majority class and *>50K* as the minority class in adult dataset. Regarding the Sex_Race label, under-represented groups include label *Female_ Amer-Indian-Eskimo* and *Female_ Other*. And over-represented groups include Sex_Race label *Female_ White* and *Male_ White*.

The following sections will show the vertical comparison among experiments without DP and horizontal comparison with fixed privacy budget and fixed data distribution scenario respectively. For every comparison, we will present the overall, per-class, and per-label performance of the model. In this chapter, we only consider the final value of metrics in the last round, specifically the accuracy, F_1-score, precision, and recall in the 500th round. Also, the range of a particular series of metric values is defined by the maximum value minus the minimum value. The standard deviation (StdDev) is calculated based on the definition of $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$.

A Vertical Comparison within Each Experiment

In this section, we only consider 6 experiments of different data distribution scenarios without introducing differential privacy. We will show the overall, per-class, and per-label model performance of theses 6 experiments. These results are regarded as the benchmark in our experiments, which will help us see the impact of different data distribution scenarios in a federated learning setting without having any privacy-preserving mechanism.

A.1 Overall and Per-class Performance

Table 1 shows the overall and per-class final accuracy of the model. As we can see from the exact values of the per-class accuracy, the accuracy on minority class is significantly lower than the majority class. Also, based on the range and standard deviation of accuracy value among different data distribution scenarios, we can see that the fluctuation of overall accuracy value is smaller than per-class accuracy values, indicating that the overall accuracy is less affected by the change of data distribution scenarios. Moreover, it also shows that under-represented class is more likely to be largely affected by the different data distribution scenarios.

Data distribution scenario	overall accuracy	<=50K accuracy	>50K accuracy
Fully IID	0.832429	0.909340	0.599475
30% 2-class nonIID	0.833948	0.918002	0.579362
50% 2-class nonIID	0.829499	0.914970	0.570617
70% 2-class nonIID	0.835793	0.927530	0.557936
Fully 2-class nonIID	0.838290	0.924065	0.578487
Normal distribution	0.834491	0.920167	0.574989
Range	0.008791	0.018190	0.041539
StdDev	0.002725	0.005921	0.012389

Table 1: Accuracy of final round among 6 experiments without DP but in different data distribution scenarios

Table 2 shows the weighted overall and per-class final F_1 score of the model. As we can see from the range and standard deviation of F_1 score among different data distribution scenarios, the fluctuation of minority class F_1 score is significantly larger than the majority class F_1 score and overall F_1 score. If we see the exact values of

the F_1 score, there is a relatively small difference between the overall F_1 score and the majority class F_1 score, but the difference between overall and minority class F_1 score is quite large.

Data distribution scenario	overall weighted F_1 score	<=50K F_1 score	>50K F_1 score
Fully IID	0.828505	0.890822	0.639757
30% 2-class nonIID	0.828418	0.892617	0.633971
50% 2-class nonIID	0.823837	0.889731	0.624253
70% 2-class nonIID	0.828422	0.894660	0.627798
Fully 2-class nonIID	0.832205	0.895746	0.639749
Normal distribution	0.828575	0.893155	0.632972
Range	0.008368	0.006015	0.015504
StdDev	0.002426	0.002067	0.005711

Table 2: F_1 score of final round among 6 experiments without DP but in different data distribution scenarios

Table 3 shows the weighted overall and per-class final precision of the model. As seen from the range and standard deviation of precision among different data distribution scenarios, the difference between overall weighted precision and majority class precision is relatively stable even when the data distribution scenario is changing. When we look at the exact value of precision, the range and standard deviation of precision in minority class are almost the same with the majority class precision, indicating that changes in data distribution scenarios does not have a significant influence difference between majority and minority subgroups of the adult dataset.

Data distribution scenario	overall weighted precision	<=50K precision	>50K precision
Fully IID	0.826578	0.873042	0.685843
30% 2-class nonIID	0.826737	0.868597	0.699947
50% 2-class nonIID	0.821956	0.865847	0.689018
70% 2-class nonIID	0.827707	0.864040	0.717660
Fully 2-class nonIID	0.830989	0.869111	0.715522
Normal distribution	0.827046	0.867683	0.703961
Range	0.009033	0.009002	0.031817
StdDev	0.002644	0.002811	0.012012

Table 3: Precision of final round among 6 experiments without DP but in different data distribution scenarios

Table 4 shows the weighted overall and per-class final recall of the model. As stated in the range of recall among different data distribution scenarios, the fluctuation of minority recall is significantly larger compared with the overall and majority class recall. Thus, it demonstrates that the minority class recall is more likely to fluctuate along with the difference in data distribution scenarios.

Data distribution scenario	overall weighted recall	<=50K recall	>50K recall
Fully IID	0.832429	0.909340	0.599475
30% 2-class nonIID	0.833948	0.918002	0.579362
50% 2-class nonIID	0.829499	0.914970	0.570617
70% 2-class nonIID	0.835793	0.927530	0.557936
Fully 2-class nonIID	0.838290	0.924065	0.578487
Normal distribution	0.834491	0.920167	0.574989
Range	0.008791	0.018190	0.041539
StdDev	0.002725	0.005921	0.012389

Table 4: Recall of final round among 6 experiments without DP but in different data distribution scenarios

A.2 Overall and Per-label Performance

Table 5 shows the overall and per-label final accuracy of the model.

Table 6 shows the overall and per-label final weighted F_1 score of the model.

Table 7 shows the overall and per-label final weighted precision of the model.

Table 8 shows the overall and per-label final weighted recall of the model.

Based on the results, the fluctuation of minority label metrics (including accuracy, F_1 score, precision, and recall) is significantly larger when the data distribution scenarios changes, compared with the range of majority label metrics. Thus, minority groups of Sex_Race labels would receive more impact when there is a change in the data distribution scenario. Also, the per-label metric values within 70% 2-class non-IID data distribution vary at most, and the per-label metric values with 30% 2-class non-IID data distribution are the most even ones among all Sex_Race labels.

B Horizontal Comparison among a series of Experiment

We have conducted 66 experiments in total, with 6 data distribution scenarios and 11 differential privacy settings. The 11 differential privacy settings include a no DP setting and 10 privacy budget values of 0.1, 0.2, 0.5, 0.8, 1, 1.2, 1.5, 2, 10, 100.

However, when ϵ is 0.1, Opacus reports an error that the privacy budget is too low to execute the whole program. Based on the differential privacy theory, this low privacy budget leads to a really high privacy level in the dataset. As a result, an extremely high privacy level would make the samples in the dataset too indistinguishable for the model to perform training. Thus, the experiment with ϵ of 0.1 will not be included in the following horizontal comparison. Both the fixed privacy budget horizontal comparison and fixed data distribution horizontal comparison will include 60 experiments, containing 6 data distribution scenarios and 10 DP settings (one no DP setting and 9 different privacy budgets).

B.1 Overall Model Performance

Table 9 shows the overall final accuracy of the model.

Table 10 shows the overall final weighted F_1 score of the model.

Table 11 shows the overall final weighted precision of the model.

Table 12 shows the overall final weighted recall of the model.

Tables above show the final overall metric values of the model. It is clear that there is no monotonic trend in the final model performance. When looking at a particular data distribution, there is no monotonic trend in the metric values when the privacy budget goes higher. Likewise, when considering a particular ϵ value for privacy budget, there is also no monotonic trend along with the change of data distribution scenarios. This phenomenon tells us that we can not simply analyze the experiment results with overall metric values only. We need to dive deeper into the per-class and per-label model performance to see the impact of various data distribution scenarios and different privacy budget settings. Since this is a binary classification task, we can investigate the model performance of the two classes separately. Especially when we conduct experiments on the highly imbalanced Adult dataset, by looking at the per-class model performance, we can see how the NN model performs in the class with more than 75% of the samples and in the class with less than 25% samples of the whole adult dataset. Besides, the

per-label model performance is also quite important for us to investigate the fairness of the model. Since the number of samples of each Sex_Race label is largely different from each other, the per-class model performance can help us see whether the change in differentially private federated learning setting would show the same impact on the majority groups and minority groups.

Moreover, since we have performed analysis of different data distributions with no DP in the vertical comparison section, we will not emphasize the no DP situations in the following horizontal comparisons.

B.2 Fixed Privacy Budget

Aside from the no DP situation and too-low privacy budget $\epsilon = 0.1$, we have 9 different privacy budgets (ϵ values), specifically with ϵ values of: 0.2, 0.5, 0.8, 1, 1.2, 1.5, 2, 10, 100.

Table 13 gives the overall, per-class, and per-label accuracy of experiments with privacy budget $\epsilon = 0.2$. As shown in the results, the overall accuracy of 6 experiments with a fixed privacy budget ϵ=0.2 does not show significant fluctuation when the data distribution scenario is changing. The significant fluctuation phenomenon happens in target class *>50K*, Sex_Race label *Female_ Amer-Indian-Eskimo*, Sex_Race label *Female_ Other*, and Sex_Race label *Male_ Other*. These are the minority classes or minority labels in the whole adult dataset. On the other side, the relatively small variation on metric values occur in target class *<=50K*, Sex_Race label *Female_ White*, and Sex_Race label *Male_ White*. These three are the majority target classes or majority labels.

Table 14 gives the overall, per-class, and per-label accuracy of experiments with privacy budget $\epsilon = 0.5$. As shown in the results, the significant fluctuation phenomenon happens in Sex_Race label *Female_ Amer-Indian-Eskimo*, *Male_ Amer-Indian-Eskimo*, *Female_ Other*, and *Male_ Other*. These are the minority labels in the whole adult dataset. On the other side, a relatively small variation on metric values occurs in target class *<=50K*, Sex_Race label *Female_ White*, and Sex_Race label *Male_ White*. These three are the majority target classes or majority labels.

Table 15 gives the overall, per-class, and per-label accuracy of experiments with privacy budget $\epsilon = 0.8$. As shown in the results, the overall accuracy of the model does not change so much when the data distribution scheme changes. The significant

fluctuation phenomenon happens in Sex_Race label *Female_ Amer-Indian-Eskimo*, *Female_ Other*, and *Male_ Other*. These are the minority labels in the whole adult dataset. On the other side, a relatively small variation on metric values occurs in Sex_Race label *Female_ White* and *Male_ White*. These three are the majority labels in the adult dataset.

Table 16 gives the overall, per-class, and per-label accuracy of experiments with privacy budget $\epsilon = 1.2$. As shown in the results, the overall accuracy of the model does not have a very large change when the data distribution scheme changes. The significant fluctuation phenomenon happens in Sex_Race label *Female_ Other*. This is a minority label in the whole adult dataset. On the other side, a relatively small variation on metric values occurs in Sex_Race label *Female_ White* and *Male_ White*. These three are the majority labels in the adult dataset.

Table 17 gives the overall, per-class, and per-label accuracy of experiments with privacy budget $\epsilon = 100$. As shown in the results, the overall accuracy of the model only shows a slightly change when the data distribution scheme changes. The significant fluctuation phenomenon happens in target class *>50K*, Sex_Race label *Female_ Other*, and Sex_Race label *Male_ Other*. These are minority labels in the whole adult dataset. On the other side, a relatively small variation on metric values occurs in target class *<=50K*, Sex_Race label *Female_ White*, and Sex_Race label *Male_ White*. These three are the majority target classes and majority labels in the adult dataset.

B.3 Fixed Data Distribution scenario

Table 18 gives the overall, per-class, and per-label accuracy of experiments with fully IID data distribution scenario. As shown in the results, the change of privacy budget (ϵ value) does not have large reflection on the overall accuracy. The target class *>50K*, Sex_Race label *Female_ Amer-Indian-Eskimo*, Sex_Race label *Female_ Other*, Sex_Race label *Male_ Amer-Indian-Eskimo*, and Sex_Race label *Male_ Other* have shown a larger fluctuation on metric values when the privacy budget ϵ changes. These are minority classes and minority labels in the whole adult dataset. On the other hand, target class *<=50K*, Sex_Race label *Female_ White*, and Sex_Race label *Male_ White* have shown a relatively small shift when ϵ changes from 0.2 to 100. And these three are the majority target classes and majority labels.

Table 19 gives the overall, per-class, and per-label accuracy of experiments with fully 2-class non-IID data distribution scenario. As shown in the results, the change of privacy budget ϵ does not have large reflection on the overall accuracy. The target class *>50K*, Sex_Race label *Female_ Amer-Indian-Eskimo*, Sex_Race label *Female_ Other*, Sex_Race label *Male_ Amer-Indian-Eskimo*, and Sex_Race label *Male_ Other* have shown a larger fluctuation on metric values when the privacy budget ϵ changes. These are minority classes and minority labels in the whole adult dataset. On the other hand, target class *<=50K*, Sex_Race label *Female_ White*, and Sex_Race label *Male_ White* have shown a relatively small shift when ϵ changes from 0.2 to 100. And these three are the majority target classes and majority labels.

Table 20 gives the overall, per-class, and per-label accuracy of experiments with normal data distribution scenario. As shown in the results, the change of privacy budget ϵ does not have large reflection on the overall accuracy. The target class *>50K*, Sex_Race label *Female_ Amer-Indian-Eskimo*, Sex_Race label *Female_ Other*, Sex_Race label *Male_ Amer-Indian-Eskimo*, and Sex_Race label *Male_ Other* have shown a larger fluctuation on metric values when the privacy budget ϵ value changes. These are minority classes and minority labels in the whole adult dataset. On the other hand, target class *<=50K*, Sex_Race label *Female_ White*, and Sex_Race label *Male_ White* have shown a relatively small shift when ϵ changes from 0.2 to 100. And these three are the majority target classes and majority labels.

Table 21 gives the overall, per-class, and per-label accuracy of experiments with 70% 2-class non-IID data distribution scenario. As shown in the results, the overall accuracy has great change along with the different values of privacy budget ϵ. The target class *>50K*, Sex_Race label *Female_ Amer-Indian-Eskimo*, and Sex_Race label *Female_ Other* have shown a larger fluctuation on he metric values when the privacy budget ϵ changes. These three are the minority classes and minority labels in the whole adult dataset. On the other hand, target class *<=50K*, Sex_Race label *Female_ White*, and Sex_Race label *Male_ White* have shown a relatively small shift when ϵ changes from 0.2 to 100. And these three are the majority target classes and majority labels.

Distribution	#samples	fully IID	30% nonIID	50% nonIID	70% nonIID	fully nonIID	normal	Range	StdDev
Female_Amer-Indian-Eskimo	166	0.794118	0.882353	0.823529	0.676471	0.911765	0.823529	0.235294	0.074825
Female_Asian-Pac-Islander	470	0.852632	0.831579	0.852632	0.821053	0.884211	0.863158	0.063158	0.020535
Female_Black	2125	0.807512	0.826291	0.814554	0.847418	0.840376	0.823944	0.039906	0.013782
Female_Other	135	0.821429	0.821429	0.892857	0.642857	0.785714	0.964286	0.321429	0.098888
Female_White	12023	0.822453	0.820790	0.831601	0.830769	0.846985	0.839917	0.026195	0.009180
Male_Amer-Indian-Eskimo	269	0.870370	0.870370	0.833333	0.833333	0.870370	0.814815	0.055555	0.022469
Male_Asian-Pac-Islander	953	0.879581	0.863874	0.801047	0.863874	0.863874	0.816754	0.078534	0.028677
Male_Black	2231	0.841163	0.816555	0.850112	0.796421	0.816555	0.821029	0.053691	0.017588
Male_Other	240	0.775510	0.897959	0.775510	0.877551	0.836735	0.857143	0.122449	0.047131
Male_White	27421	0.836463	0.839562	0.828806	0.841203	0.833911	0.833546	0.012397	0.004104
Range		0.104071	0.0081404	0.117347	0.234694	0.126051	0.149471		
StdDev		0.031079	0.027690	0.029948	0.075108	0.033722	0.042479		

Table 5: Accuracy score of final round per Sex_Race label among 6 experiments without DP but in different data distribution scenarios

Distribution	#samples	Range	fully IID	30% nonIID	50% nonIID	70% nonIID	fully nonIID	normal
Female_Amer-Indian-Eskimo	166	0.112188	0.798155	0.882353	0.811975	0.647964	0.910343	0.808258
Female_Asian-Pac-Islander	470	0.040803	0.849495	0.822963	0.846358	0.816307	0.883323	0.857110
Female_Black	2125	0.037531	0.802986	0.819560	0.809293	0.840517	0.839795	0.818370
Female_Other	135	0.371222	0.820252	0.807053	0.890592	0.591837	0.741071	0.963059
Female_White	12023	0.025742	0.817943	0.815490	0.827528	0.825393	0.841232	0.835130
Male_Amer-Indian-Eskimo	269	0.083624	0.868956	0.881243	0.813402	0.835152	0.863298	0.797619
Male_Asian-Pac-Islander	953	0.087126	0.881045	0.859539	0.793919	0.861547	0.853681	0.814723
Male_Black	2231	0.060973	0.835326	0.806293	0.844017	0.783044	0.813185	0.809564
Male_Other	240	0.147289	0.772934	0.895588	0.748299	0.872303	0.832138	0.845667
Male_White	27421	0.011597	0.832622	0.834293	0.822696	0.833212	0.827068	0.827594
Range			0.108111	0.288535	0.142298	0.280466	0.169272	0.165440

Table 6: Weighted F_1 score of final round per Sex_Race label among 6 experiments without DP but in different data distribution scenarios

Distribution	#samples	Range	fully IID	30% nonIID	50% nonIID	70% nonIID	fully nonIID	normal
Female_Amer-Indian-Eskimo	166	0.283669	0.803660	0.882353	0.808114	0.627181	0.910850	0.824930
Female_Asian-Pac-Islander	470	0.069110	0.847360	0.823058	0.844270	0.813573	0.882683	0.857614
Female_Black	2125	0.038744	0.800516	0.817452	0.807878	0.839204	0.839260	0.816810
Female_Other	135	0.358696	0.821238	0.818634	0.891156	0.607143	0.741758	0.965839
Female_White	12023	0.026269	0.816026	0.813701	0.825652	0.823404	0.839970	0.833148
Male_Amer-Indian-Eskimo	269	0.082793	0.868254	0.898765	0.815972	0.837607	0.875882	0.830303
Male_Asian-Pac-Islander	953	0.091015	0.883043	0.859166	0.792028	0.860196	0.854194	0.813920
Male_Black	2231	0.060870	0.832712	0.809547	0.842943	0.782073	0.811199	0.812219
Male_Other	240	0.122449	0.775510	0.897959	0.775510	0.877551	0.836735	0.857143
Male_White	27421	0.012933	0.830746	0.832543	0.820760	0.833693	0.826246	0.825938
Range			0.107583	0.088412	0.115646	0.270408	0.169092	0.153620

Table 7: Weighted precision of final round per Sex_Race label among 6 experiments without DP but in different data distribution scenarios

Distribution	#samples	Range	fully IID	30% nonIID	50% nonIID	70% nonIID	fully nonIID	normal
Female_Amer-Indian-Eskimo	166	0.235294	0.794118	0.882353	0.823529	0.676471	0.911765	0.823529
Female_Asian-Pac-Islander	470	0.063158	0.852632	0.831579	0.852632	0.821053	0.884211	0.863158
Female_Black	2125	0.039906	0.807512	0.826291	0.814554	0.847418	0.840376	0.823944
Female_Other	135	0.321429	0.821429	0.821429	0.892857	0.642857	0.785714	0.964286
Female_White	12023	0.026195	0.822453	0.820790	0.831601	0.830769	0.846985	0.839917
Male_Amer-Indian-Eskimo	269	0.055555	0.870370	0.870370	0.833333	0.833333	0.870370	0.814815
Male_Asian-Pac-Islander	953	0.078534	0.879581	0.863874	0.801047	0.863874	0.863874	0.816754
Male_Black	2231	0.053691	0.841163	0.816555	0.850112	0.796421	0.816555	0.821029
Male_Other	240	0.112449	0.775510	0.897959	0.775510	0.877551	0.836735	0.857143
Male_White	27421	0.010756	0.836463	0.839562	0.828806	0.841203	0.833911	0.833546
Range			0.104071	0.081404	0.117347	0.234694	0.126051	0.149471

Table 8: Weighted recall of final round per Sex_Race label among 6 experiments without DP but in different data distribution scenarios

	no DP	ϵ=0.2	ϵ=0.5	ϵ=0.8	ϵ=1	ϵ=1.2	ϵ=1.5	ϵ=2	ϵ=10	ϵ=100	Range	StdDev
fully IID	0.832429	0.759171	0.803343	0.818320	0.820056	0.827871	0.826026	0.828196	0.827437	0.831452	0.073258	0.021026
30% nonIID	0.833948	0.776210	0.823204	0.826243	0.825157	0.833623	0.832972	0.826894	0.824072	0.829824	0.057738	0.016122
50% nonIID	0.829499	0.813436	0.823855	0.830041	0.830475	0.834382	0.832212	0.828522	0.829607	0.833623	0.020946	0.005758
70% nonIID	0.835793	0.790211	0.822770	0.829390	0.827979	0.829933	0.826134	0.830041	0.836445	0.829499	0.046234	0.012472
fully nonIID	0.838290	0.783916	0.828305	0.828088	0.827545	0.828739	0.832646	0.830475	0.829064	0.829499	0.054374	0.014235
normal distribution	0.834491	0.782939	0.828413	0.827979	0.827871	0.828956	0.829607	0.827871	0.834274	0.826243	0.051552	0.014214
Range	0.008791	0.054265	0.025070	0.011721	0.010419	0.006511	0.006946	0.003581	0.012373	0.007380		
StdDev	0.002725	0.016242	0.008502	0.003924	0.003273	0.002500	0.002932	0.001237	0.003321	0.002232		

Table 9: Overall accuracy of final round among 60 experiments with 6 data distribution scenarios and 10 DP settings

	no DP	ϵ=0.2	ϵ=0.5	ϵ=0.8	ϵ=1	ϵ=1.2	ϵ=1.5	ϵ=2	ϵ=10	ϵ=100
fully IID	0.828505	0.745091	0.788369	0.807595	0.804865	0.816664	0.819427	0.818812	0.817091	0.822951
30% nonIID	0.828418	0.757198	0.814171	0.814199	0.815095	0.826154	0.824291	0.818012	0.814344	0.820585
50% nonIID	0.823837	0.805062	0.814390	0.821302	0.820623	0.825721	0.823949	0.818905	0.820659	0.825874
70% nonIID	0.828422	0.774157	0.809742	0.818605	0.820004	0.819990	0.817572	0.821078	0.828965	0.820899
fully nonIID	0.832205	0.782696	0.819514	0.819210	0.819912	0.820268	0.824259	0.821498	0.820591	0.820750
normal distribution	0.828575	0.785841	0.817587	0.819040	0.820413	0.821260	0.819509	0.819170	0.826308	0.817365

Table 10: Overall weighted F_1 score of final round among 60 experiments with 6 data distribution scenarios and 10 DP settings

	no DP	ϵ=0.2	ϵ=0.5	ϵ=0.8	ϵ=1	ϵ=1.2	ϵ=1.5	ϵ=2	ϵ=10	ϵ=100
fully IID	0.826578	0.739192	0.788592	0.807150	0.808967	0.818051	0.817601	0.818672	0.817621	0.822550
30% nonIID	0.826737	0.754611	0.813222	0.816138	0.815078	0.825300	0.824195	0.817368	0.813944	0.820537
50% nonIID	0.821956	0.802915	0.813785	0.820911	0.821163	0.825777	0.823465	0.818985	0.820367	0.825193
70% nonIID	0.827707	0.772690	0.812029	0.819835	0.818923	0.820532	0.816648	0.820849	0.828381	0.820352
fully nonIID	0.830989	0.781591	0.818965	0.818696	0.818614	0.819556	0.823909	0.821330	0.819915	0.820305
normal distribution	0.827046	0.789495	0.818699	0.818555	0.819058	0.820117	0.820140	0.818514	0.825836	0.816647

Table 11: Overall weighted precision of final round among 60 experiments with 6 data distribution scenarios and 10 DP settings

	no DP	ϵ=0.2	ϵ=0.5	ϵ=0.8	ϵ=1	ϵ=1.2	ϵ=1.5	ϵ=2	ϵ=10	ϵ=100
fully IID	0.832429	0.759171	0.803343	0.818320	0.820056	0.827871	0.826026	0.828196	0.827437	0.831452
30% nonIID	0.833948	0.776210	0.823204	0.826243	0.825157	0.833623	0.832972	0.826894	0.824072	0.829824
50% nonIID	0.829499	0.813436	0.823855	0.830041	0.830475	0.834382	0.832212	0.828522	0.829607	0.833623
70% nonIID	0.835793	0.790211	0.822770	0.829390	0.827979	0.829933	0.826134	0.830041	0.836445	0.829499
fully nonIID	0.838290	0.783916	0.828305	0.828088	0.827545	0.828739	0.832646	0.830475	0.829064	0.829499
normal distribution	0.834491	0.782939	0.828413	0.827979	0.827871	0.828956	0.829607	0.827871	0.834274	0.826243

Table 12: Overall weighted recall of final round among 60 experiments with 6 data distribution scenarios and 10 DP settings

	#samples	fully IID	30% nonIID	50% nonIID	70% nonIID	fully nonIID	normal	Range	StdDev
Overall	46033	0.759171	0.776210	0.813436	0.790211	0.783916	0.782939	0.054265	0.016242
<= 50K	34611	0.885520	0.912372	0.912661	0.916991	0.861701	0.841490	0.075501	0.028584
>50K	11422	0.376476	0.363795	0.512899	0.406209	0.548317	0.605597	0.241802	0.091695
Female_Amer-Indian-Eskimo	166	0.676471	0.647059	0.882353	0.764706	0.882353	0.794118	0.235294	0.090918
Female_Asian-Pac-Islander	470	0.810526	0.757895	0.810526	0.810526	0.694737	0.768421	0.115789	0.041959
Female_Black	2125	0.739437	0.784038	0.821596	0.795775	0.781690	0.795775	0.082159	0.024657
Female_Other	135	0.821429	0.892857	0.857143	0.785714	0.821429	0.750000	0.142857	0.046107
Female_White	12023	0.765073	0.768815	0.808316	0.793347	0.783368	0.781289	0.043243	0.014565
Male_Amer-Indian-Eskimo	209	0.685185	0.685185	0.740741	0.685185	0.796296	0.740741	0.111111	0.041409
Male_Asian-Pac-Islander	953	0.806283	0.753927	0.832461	0.811518	0.811518	0.801047	0.078534	0.023929
Male_Black	2231	0.780761	0.762864	0.816555	0.762864	0.791946	0.776286	0.053691	0.018527
Male_Other	240	0.755102	0.775510	0.653061	0.734694	0.714286	0.755102	0.122449	0.039812
Male_White	27421	0.754786	0.782133	0.815679	0.791249	0.783956	0.783592	0.060893	0.017814

Table 13: Overall, per-class, and per-label final accuracy with 6 data distribution scenarios and a fixed privacy budget $\epsilon = 0.2$

	#samples	fully IID	30% nonIID	50% nonIID	70% nonIID	fully nonIID	normal	Range
Overall	46033	0.803343	0.823204	0.823855	0.822770	0.828305	0.828413	0.025070
<= 50K	34611	0.925509	0.923343	0.925509	0.936914	0.926808	0.934459	0.013571
>50K	11422	0.433319	0.519895	0.515960	0.477044	0.529952	0.507215	0.096633
Female_Amer-Indian-Eskimo	166	0.852941	0.794118	0.882353	0.823529	0.794118	0.764706	0.117647
Female_Asian-Pac-Islander	470	0.842105	0.831579	0.884211	0.800000	0.821053	0.810526	0.084211
Female_Black	2125	0.798122	0.814554	0.835681	0.823944	0.807512	0.823944	0.037559
Female_Other	135	0.678571	0.964286	0.642857	0.821429	0.642857	0.892857	0.321429
Female_White	12023	0.802911	0.827027	0.835343	0.823701	0.834511	0.831185	0.032432
Male_Amer-Indian-Eskimo	209	0.740741	0.703704	0.777778	0.851852	0.814815	0.851852	0.148148
Male_Asian-Pac-Islander	953	0.816754	0.827225	0.863874	0.806283	0.863874	0.827225	0.057591
Male_Black	2231	0.800895	0.823266	0.832215	0.812081	0.803132	0.814318	0.031320
Male_Other	240	0.918367	0.795918	0.897959	0.795918	0.857143	0.734694	0.183673
Male_White	27421	0.802917	0.822789	0.815132	0.824066	0.829170	0.829717	0.026800

Table 14: Overall, per-class, and per-label final accuracy with 6 data distribution scenarios and a fixed privacy budget $\epsilon = 0.5$

	#samples	fully IID	30% nonIID	50% nonIID	70% nonIID	fully nonIID	normal	Range
Overall	46033	0.818320	0.826243	0.830041	0.829390	0.828088	0.827979	0.011721
<= 50K	34611	0.925220	0.936769	0.928107	0.935181	0.926953	0.927097	0.011549
>50K	11422	0.494534	0.491474	0.533013	0.508964	0.528640	0.527766	0.041539
Female_Amer-Indian-Eskimo	166	0.823529	0.882353	0.735294	0.823529	0.823529	0.705882	0.176471
Female_Asian-Pac-Islander	470	0.831579	0.800000	0.821053	0.884211	0.831579	0.863158	0.084211
Female_Black	2125	0.835681	0.798122	0.826291	0.809859	0.807512	0.814554	0.037559
Female_Other	135	0.821429	0.857143	0.785714	0.750000	0.821429	0.678571	0.178572
Female_White	12023	0.819958	0.821206	0.827027	0.828274	0.826195	0.832017	0.012059
Male_Amer-Indian-Eskimo	209	0.777778	0.814815	0.814815	0.740741	0.814815	0.814815	0.074074
Male_Asian-Pac-Islander	953	0.858639	0.848168	0.848168	0.863874	0.848168	0.790576	0.073298
Male_Black	2231	0.803132	0.821029	0.841163	0.814318	0.845638	0.832215	0.042506
Male_Other	240	0.816327	0.795918	0.795918	0.816327	0.693878	0.734694	0.122449
Male_White	27421	0.816226	0.830629	0.831541	0.831905	0.829717	0.830082	0.015679

Table 15: Overall, per-class, and per-label final accuracy with 6 data distribution scenarios and a fixed privacy budget $\epsilon = 0.8$

	#samples	fully IID	30% nonIID	50% nonIID	70% nonIID	fully nonIID	normal	Range	StdDev
Overall	46033	0.827871	0.833623	0.834382	0.829933	0.828739	0.828956	0.006511	0.002500
<= 50K	34611	0.935326	0.926086	0.931572	0.932583	0.925942	0.923055	0.012271	0.004332
>50K	11422	0.502405	0.553564	0.540009	0.519021	0.534324	0.543944	0.051159	0.016937
Female_Amer-Indian-Eskimo	166	0.735294	0.823529	0.882353	0.764706	0.882353	0.705882	0.176471	0.068802
Female_Asian-Pac-Islander	470	0.831579	0.810526	0.842105	0.778947	0.936842	0.789474	0.157895	0.051925
Female_Black	2125	0.880282	0.847418	0.812207	0.816901	0.821596	0.828638	0.068075	0.023347
Female_Other	135	0.892857	0.821429	0.892857	0.678571	0.785714	0.928571	0.250000	0.084179
Female_White	12023	0.819543	0.832017	0.828690	0.818711	0.814553	0.829106	0.017464	0.006443
Male_Amer-Indian-Eskimo	209	0.870370	0.777778	0.870370	0.814815	0.870370	0.796296	0.092592	0.038549
Male_Asian-Pac-Islander	953	0.821990	0.774869	0.811518	0.827225	0.837696	0.853403	0.078534	0.024448
Male_Black	2231	0.812081	0.841163	0.841163	0.838926	0.838926	0.832215	0.029082	0.010286
Male_Other	240	0.836735	0.857143	0.795918	0.857143	0.857143	0.795918	0.061225	0.027423
Male_White	27421	0.828624	0.835552	0.838104	0.837192	0.831723	0.829353	0.009480	0.003722

Table 16: Overall, per-class, and per-label final accuracy with 6 data distribution scenarios and a fixed privacy budget $\epsilon = 1.2$

	#samples	fully IID	30% nonIID	50% nonIID	70% nonIID	fully nonIID	normal	Range	StdDev
Overall	46033	0.831452	0.829824	0.833623	0.829499	0.829499	0.826243	0.007380	0.002232
<= 50K	34611	0.928396	0.929840	0.927241	0.927097	0.927674	0.925365	0.004475	0.001356
>50K	11422	0.537822	0.526891	0.550066	0.533887	0.532138	0.526017	0.024049	0.008053
Female_Amer-Indian-Eskimo	166	0.852941	0.823529	0.764706	0.911765	0.735294	0.823529	0.176471	0.057376
Female_Asian-Pac-Islander	470	0.873684	0.800000	0.821053	0.831579	0.852632	0.757895	0.115789	0.037175
Female_Black	2125	0.854460	0.852113	0.830986	0.826291	0.823944	0.833333	0.030516	0.012027
Female_Other	135	0.642857	0.785714	0.928571	0.785714	0.892857	0.785714	0.285714	0.091636
Female_White	12023	0.820374	0.824948	0.832432	0.829938	0.834096	0.824532	0.013722	0.004823
Male_Amer-Indian-Eskimo	209	0.851852	0.851852	0.870370	0.814815	0.796296	0.759259	0.111111	0.038177
Male_Asian-Pac-Islander	953	0.832461	0.790576	0.785340	0.842932	0.874346	0.827225	0.089006	0.030479
Male_Black	2231	0.845638	0.827740	0.836689	0.823266	0.823266	0.852349	0.029083	0.011142
Male_Other	240	0.816327	0.795918	0.938776	0.816327	0.795918	0.877551	0.142858	0.051920
Male_White	27421	0.833364	0.832634	0.834640	0.829535	0.827347	0.825889	0.008751	0.003215

Table 17: Overall, per-class, and per-label final accuracy with 6 data distribution scenarios and a fixed privacy budget $\epsilon = 100$

	#samples	no DP	ϵ=0.2	ϵ=0.5	ϵ=0.8	ϵ=1	ϵ=1.2	ϵ=1.5	ϵ=2	ϵ=10	ϵ=100	Range	StdDev
Overall	46033	0.832429	0.759171	0.803343	0.818320	0.820056	0.827871	0.826026	0.828196	0.827437	0.831452	0.073258	0.021026
<= 50K	34611	0.909340	0.885520	0.925509	0.925220	0.941244	0.935326	0.916125	0.928974	0.931861	0.928396	0.055724	0.015091
>50K	11422	0.599475	0.376476	0.433319	0.494534	0.452995	0.502405	0.553126	0.522956	0.511150	0.537822	0.222999	0.060601
Female_Amer-Indian-Eskimo	166	0.794118	0.676471	0.852941	0.823529	0.911765	0.735294	0.911765	0.794118	0.941176	0.852941	0.264705	0.078701
Female_Asian-Pac-Islander	470	0.852632	0.810526	0.842105	0.831579	0.873684	0.831579	0.894737	0.821053	0.842105	0.873684	0.084211	0.025021
Female_Black	2125	0.807512	0.739437	0.798122	0.835681	0.821596	0.880282	0.838028	0.800469	0.798122	0.854460	0.140845	0.036484
Female_Other	135	0.821429	0.821429	0.678571	0.821429	0.857143	0.892857	0.785714	0.928571	0.821429	0.642857	0.285714	0.083299
Female_White	12023	0.822453	0.765073	0.802911	0.819958	0.822869	0.819543	0.825780	0.829106	0.833680	0.820374	0.068607	0.018657
Male_Amer-Indian-Eskimo	269	0.870370	0.685185	0.740741	0.777778	0.833333	0.870370	0.814815	0.777778	0.833333	0.851852	0.185185	0.056928
Male_Asian-Pac-Islander	953	0.879581	0.806283	0.816754	0.858639	0.790576	0.821990	0.837696	0.848168	0.848168	0.832461	0.089005	0.024895
Male_Black	2231	0.841163	0.780761	0.800895	0.803132	0.800895	0.812081	0.834452	0.823266	0.841163	0.845638	0.064877	0.020886
Male_Other	240	0.775510	0.755102	0.918367	0.816327	0.734694	0.836735	0.795918	0.959184	0.816327	0.816327	0.224490	0.065846
Male_White	27421	0.836463	0.754786	0.802917	0.816226	0.820237	0.828624	0.822972	0.828806	0.824248	0.833364	0.081677	0.022559

Table 18: Overall, per-class, and per-label final accuracy with 10 DP settings and a fixed fully IID data distribution scenario

	#samples	no DP	ϵ=0.2	ϵ=0.5	ϵ=0.8	ϵ=1	ϵ=1.2	ϵ=1.5	ϵ=2	ϵ=10	ϵ=100	Range	StdDev
Overall	46033	0.83829	0.783916	0.828305	0.828088	0.827545	0.828739	0.832646	0.830475	0.829064	0.829499	0.054374	0.014235
<= 50K	34611	00.924065	0.861701	0.926808	0.926953	0.921611	0.925942	0.928974	0.929407	0.926231	0.927674	0.067706	0.019531
>50K	11422	0.578487	0.548317	0.529952	0.52864	0.542632	0.534324	0.540883	0.530826	0.534762	0.532138	0.049847	0.014120
Female_Amer-Indian-Eskimo	166	0.911765	0.882353	0.794118	0.823529	0.882353	0.882353	0.882353	0.970588	0.823529	0.735294	0.235294	0.062806
Female_Asian-Pac-Islander	470	0.884211	0.694737	0.821053	0.831579	0.873684	0.936842	0.778947	0.810526	0.842105	0.852632	0.242105	0.061819
Female_Black	2125	0.840376	0.78169	0.807512	0.807512	0.840376	0.821596	0.816901	0.816901	0.821596	0.823944	0.058686	0.016134
Female_Other	135	0.785714	0.821429	0.642857	0.821429	0.892857	0.785714	0.892857	0.857143	0.785714	0.892857	0.250000	0.072228
Female_White	12023	0.846985	0.783368	0.834511	0.826195	0.827859	0.814553	0.821622	0.827443	0.826195	0.834096	0.063617	0.015881
Male_Amer-Indian-Eskimo	269	0.87037	0.796296	0.814815	0.814815	0.833333	0.87037	0.87037	0.777778	0.796296	0.796296	0.092592	0.033385
Male_Asian-Pac-Islander	953	0.863874	0.811518	0.863874	0.848168	0.82199	0.837696	0.853403	0.853403	0.842932	0.874346	0.062828	0.018407
Male_Black	2231	0.816555	0.791946	0.803132	0.845638	0.850112	0.838926	0.845638	0.814318	0.805369	0.823266	0.058166	0.019489
Male_Other	240	0.836735	0.714286	0.857143	0.693878	0.857143	0.857143	0.918367	0.857143	0.795918	0.795918	0.224489	0.066099
Male_White	27421	0.833911	0.783956	0.82917	0.829717	0.822972	0.831723	0.836098	0.832999	0.832999	0.827347	0.052142	0.014484

Table 19: Overall, per-class, and per-label final accuracy with 10 DP settings and a fixed fully 2-class non-IID data distribution scenario

	#samples	no DP	ϵ=0.2	ϵ=0.5	ϵ=0.8	ϵ=1	ϵ=1.2	ϵ=1.5	ϵ=2	ϵ=10	ϵ=100	Range	StdDev
Overall	46033	0.834491	0.782939	0.828413	0.827979	0.827871	0.828956	0.829607	0.827871	0.834274	0.826243	0.051552	0.014214
<= 50K	34611	0.920167	0.84149	0.934459	0.927097	0.921178	0.923055	0.932871	0.926086	0.928685	0.925365	0.092969	0.025890
>50K	11422	0.574989	0.605597	0.507215	0.527766	0.545256	0.543944	0.516834	0.530389	0.548317	0.526017	0.098382	0.027603
Female_Amer-Indian-Eskimo	166	0.823529	0.794118	0.764706	0.705882	0.794118	0.705882	0.823529	0.823529	0.852941	0.823529	0.147059	0.048239
Female_Asian-Pac-Islander	470	0.863158	0.768421	0.810526	0.863158	0.821053	0.789474	0.884211	0.863158	0.863158	0.757895	0.126316	0.042900
Female_Black	2125	0.823944	0.795775	0.823944	0.814554	0.826291	0.828638	0.852113	0.828638	0.842723	0.833333	0.056338	0.014396
Female_Other	135	0.964286	0.75	0.892857	0.678571	0.785714	0.928571	0.928571	0.821429	0.892857	0.785714	0.285715	0.087773
Female_White	12023	0.839917	0.781289	0.831185	0.832017	0.828274	0.829106	0.829106	0.821206	0.828274	0.824532	0.058628	0.015126
Male_Amer-Indian-Eskimo	269	0.814815	0.740741	0.851852	0.814815	0.796296	0.796296	0.796296	0.851852	0.796296	0.759259	0.111111	0.033179
Male_Asian-Pac-Islander	953	0.816754	0.801047	0.827225	0.790576	0.837696	0.853403	0.806283	0.811518	0.848168	0.827225	0.062827	0.019449
Male_Black	2231	0.821029	0.776286	0.814318	0.832215	0.836689	0.832215	0.803132	0.845638	0.841163	0.852349	0.076063	0.021574
Male_Other	240	0.857143	0.755102	0.734694	0.734694	0.918367	0.795918	0.877551	0.755102	0.836735	0.877551	0.183673	0.064180
Male_White	27421	0.833546	0.783592	0.829717	0.830082	0.8268	0.829353	0.829535	0.829717	0.83464	0.825889	0.051048	0.014120

Table 20: Overall, per-class, and per-label final accuracy with 10 DP settings and a fixed normal data distribution scenario

	#samples	no DP	ϵ=0.2	ϵ=0.5	ϵ=0.8	ϵ=1	ϵ=1.2	ϵ=1.5	ϵ=2	ϵ=10	ϵ=100	Range
Overall	46033	0.835793	0.790211	0.822770	0.829390	0.827979	0.829933	0.826134	0.830041	0.836445	0.829499	0.046234
<= 50K	34611	0.927530	0.916991	0.936914	0.935181	0.923343	0.932583	0.924065	0.928974	0.928540	0.927097	0.010077
>50K	11422	0.557936	0.406209	0.477044	0.508964	0.539134	0.519021	0.529515	0.530389	0.557499	0.533887	0.151727
Female_Amer-Indian-Eskimo	166	0.676471	0.764706	0.823529	0.823529	0.764706	0.764706	0.705882	0.764706	0.823529	0.911765	0.235294
Female_Asian-Pac-Islander	470	0.821053	0.810526	0.800000	0.884211	0.852632	0.778947	0.831579	0.789474	0.852632	0.831579	0.105264
Female_Black	2125	0.847418	0.795775	0.823944	0.809859	0.823944	0.816901	0.835681	0.795775	0.863850	0.826291	0.068075
Female_Other	135	0.642857	0.785714	0.821429	0.750000	0.928571	0.678571	0.928571	0.785714	0.821429	0.785714	0.285714
Female_White	12023	0.830769	0.793347	0.823701	0.828274	0.826611	0.818711	0.828690	0.832432	0.829106	0.829938	0.039085
Male_Amer-Indian-Eskimo	269	0.833333	0.685185	0.851852	0.740741	0.870370	0.814815	0.833333	0.870370	0.907407	0.814815	0.166666
Male_Asian-Pac-Islander	953	0.863874	0.811518	0.806283	0.863874	0.801047	0.827225	0.790576	0.821990	0.853403	0.842932	0.073298
Male_Black	2231	0.796421	0.762864	0.812081	0.814318	0.847875	0.838926	0.794183	0.812081	0.836689	0.823266	0.085011
Male_Other	240	0.877551	0.734694	0.795918	0.816327	0.877551	0.857143	0.795918	0.897959	0.836735	0.816327	0.163265
Male_White	27421	0.841203	0.791249	0.824066	0.831905	0.826800	0.837192	0.828441	0.833728	0.836098	0.829535	0.049954

Table 21: Overall, per-class, and per-label final accuracy with 10 DP settings and a fixed 70% 2-class non-IID data distribution scenario