Constrained Differentially Private Federated Learning for Low-bandwidth Devices

Raouf Kerkouche, Gergely Ács, Claude Castelluccia, Pierre Genevès

Constrained Differentially Private Federated Learning for Low-bandwidth Devices

Raouf Kerkouche[1] **Gergely Ács**[2] **Claude Castelluccia**[1] **Pierre Genevès**[3]

[1]Privatics team, Univ. Grenoble Alpes, Inria, 38000 Grenoble, France,
[2]Crysys Lab, BME-HIT
[3] Tyrex team, Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG

Abstract

Federated learning becomes a prominent approach when different entities want to learn collaboratively a common model without sharing their training data. However, Federated learning has two main drawbacks. First, it is quite bandwidth inefficient as it involves a lot of message exchanges between the aggregating server and the participating entities. This bandwidth and corresponding processing costs could be prohibitive if the participating entities are, for example, mobile devices. Furthermore, although federated learning improves privacy by not sharing data, recent attacks have shown that it still leaks information about the training data.
This paper presents a novel privacy-preserving federated learning scheme. The proposed scheme provides theoretical privacy guarantees, as it is based on Differential Privacy. Furthermore, it optimizes the model accuracy by constraining the model learning phase on few selected weights. Finally, as shown experimentally, it reduces the upstream *and* downstream bandwidth by up to 99.9% compared to standard federated learning, making it practical for mobile systems.

1 INTRODUCTION

In Machine Learning, different entities may want to collaborate in order to improve their local model accuracy. In traditional machine learning, such collaboration requires to first store all entities' data on a centralized server and then to train a model on it. Such data centralization might be problematic when the data are sensitive and data privacy is required. In order to mitigate this problem, Federated learning, which allows different entities to learn collaboratively a common model without sharing their data, was introduced [Shokri and Shmatikov, 2015, McMahan et al., 2016]. Instead of sharing the training data, Federated Learning shares the model parameters between a server, which plays the role of aggregator, and the participating entities. Although Federated Learning improves privacy, model parameters can leak information about the training data. Indeed, Zhu et al. [2019], Zhao et al. [2020], Geiping et al. [2020] presented some attacks that allow an adversary to reconstruct pieces of the training data of some entities. Nasr et al. [2019] define a membership attack that allows to infer if a particular record is included in the data of a specific entity. Similarly, Melis et al. [2018] define an attack which aims at inferring if a subgroup of people with a specific property, like for example skin color or ethnicity, is included in the dataset of a particular participating entity. A solution to prevent these attacks and provide theoretical guarantees in to use a privacy model called Differential Privacy [Dwork and Roth, 2014]. Differential Privacy has been applied to federated learning in order to protect either each record included in the dataset of any entity (record-level guarantee), or the whole dataset of any entity (client-level guarantee). Unfortunately, it is well-known that Differential Privacy drastically degrades the accuracy of the global model as it requires to add random noise to the gradients (record-level) or to the updates (client-level) of each client. Recent work by Kerkouche et al. [2020] shows that this accuracy penalty can be reduced if the model is compressed, as compression reduces the required amount of noise. Furthermore, Kerkouche et al. [2020] show that accuracy can be further improved by adding noise only to the largest update's values as adding noise on values close to 0 is likely to lead to random update values.

Following up on these results, we propose a novel differentially private federated learning solution that improves the model accuracy (1) by updating only a fixed subset of the model weights, and (2) by maintaining the other weights constant. The proposed scheme provides theoretical privacy guarantees, as it is based on Differential Privacy. Furthermore, it optimizes the model accuracy by constraining the model learning phase on a few selected weights. As all participants always update the same set of weights and transfer

them to the server for aggregation, the proposal can be easily integrated with secure aggregation [Bonawitz et al., 2016], which allows parties to add less noise than other decentralized perturbation approaches such as randomized response [Erlingsson et al., 2014] used in local differential privacy. Moreover, it also reduces the upstream and downstream bandwidth by a factor of 1000 compared to standard federated learning, making it practical for mobile systems. The paper is structured as follows: In Section 2 we introduce the necessary background to understand the proposal, in Section 3 we define our solution called FL-TOP and in Section 3.2 its private extension called FL-TOP-DP.

2 BACKGROUND

2.1 FEDERATED LEARNING (FL-STD)

In federated learning [Shokri and Shmatikov, 2015, McMahan et al., 2016], multiple parties (clients) build a common machine learning model from union of their training data without sharing them with each other. At each round of the training, a selected set of clients retrieve the global model from the parameter server, update the global model based on their own training data, and send back their updated model to the server. The server aggregates the updated models of all clients to obtain a global model that is re-distributed to some selected parties in the next round.

In particular, a subset $\mathbb{K}$ of all N clients are randomly selected at each round to update the global model, and $C = |\mathbb{K}|/N$ denotes the fraction of selected clients. At round t, a selected client $k \in \mathbb{K}$ executes T_{gd} local gradient descent iterations on the common model $\mathbf{w}_{t-1}$ using its own training data D_k ($D = \cup_{k\in\mathbb{K}} D_k$), and obtains the updated model $\mathbf{w}_t^k$, where the number of weights is denoted by n (i.e., $|\mathbf{w}_t^k| = |\Delta\mathbf{w}_t^k| = n$ for all k and t). Each client k submits the update $\Delta\mathbf{w}_t^k = \mathbf{w}_t^k - \mathbf{w}_{t-1}^k$ to the server, which then updates the common model as follows: $\mathbf{w}_t = \mathbf{w}_{t-1} + \sum_{k\in\mathbb{K}} \frac{|D_k|}{\sum_j |D_j|} \Delta\mathbf{w}_t^k$, where $|D_k|$ is known to the server for all k (a client's update is weighted with the size of its training data). The server stops training after a fixed number of rounds T_{cl}, or when the performance of the common model does not improve on a held-out data.

Note that each D_k may be generated from different distributions (i.e., Non-IID case), that is, any client's local dataset may not be representative of the population distribution [McMahan et al., 2016]. This can happen, for example, when not all output classes are represented in every client's training data. The federated learning of neural networks is summarized in Alg. 4. In the sequel, each client is assumed to use the same model architecture.

The motivation of federated learning is three-fold: first, it aims to provide confidentiality of each participant's training data by sharing only model updates instead of potentially sensitive training data. Second, in order to decrease communication costs, clients can perform multiple local SGD iterations before sending their update back to the server. Third, in each round, only a few clients are required to perform local training of the common model, which further diminishes communication costs and makes the approach especially appealing with large number of clients.

However, several prior works have demonstrated that model updates do leak potentially sensitive information [Nasr et al., 2019, Melis et al., 2018]. Hence, simply not sharing training data *per se* is not enough to guarantee their confidentiality.

2.2 DIFFERENTIAL PRIVACY

Differential privacy allows a party to privately release information about a dataset: a function of an input dataset is perturbed, so that any information which can differentiate a record from the rest of the dataset is bounded Dwork and Roth [2014].

Definition 1 (Privacy loss)**.** *Let $\mathscr{A}$ be a privacy mechanism which assigns a value Range($\mathscr{A}$) to a dataset D. The privacy loss of $\mathscr{A}$ with datasets D and D′ at output $O \in$ Range($\mathscr{A}$) is a random variable $\mathscr{P}(\mathscr{A}, D, D', O) = \log \frac{\Pr[\mathscr{A}(D)=O]}{\Pr[\mathscr{A}(D')=O]}$ where the probability is taken on the randomness of $\mathscr{A}$.*

Definition 2 ((ε,δ)-Differential Privacy [Dwork and Roth, 2014])**.** *A privacy mechanism $\mathscr{A}$ guarantees (ε,δ)-differential privacy if for any database D and D′, differing on at most one record, $\Pr_{O\sim\mathscr{A}(D)}[\mathscr{P}(\mathscr{A}, D, D', O) > \varepsilon] \leq \delta$.*

Intuitively, this guarantees that an adversary, provided with the output of $\mathscr{A}$, can draw almost the same conclusions (up to ε with probability larger than $1-\delta$) about any record no matter if it is included in the input of $\mathscr{A}$ or not. That is, for any record owner, a privacy breach is unlikely to be due to its participation in the dataset.

Moments Accountant. Differential privacy maintains composition; the privacy guarantee of the k-fold adaptive composition of $\mathscr{A}_{1:k} = \mathscr{A}_1, \ldots, \mathscr{A}_k$ can be computed using the moments accountant method Abadi et al. [2016]. In particular, it follows from Markov's inequality that $\Pr[\mathscr{P}(\mathscr{A}, D, D', O) \geq \varepsilon] \leq \mathbb{E}[\exp(\lambda\mathscr{P}(\mathscr{A}, D, D', O))]/\exp(\lambda\varepsilon)$ for any output $O \in$ Range($\mathscr{A}$) and $\lambda > 0$. $\mathscr{A}$ is (ε,δ)-DP with $\delta = \min_\lambda \exp(\alpha_\mathscr{A}(\lambda) - \lambda\varepsilon)$, where $\alpha_\mathscr{A}(\lambda) = \max_{D,D'} \log \mathbb{E}_{O\sim\mathscr{A}(D)}[\exp(\lambda\mathscr{P}(\mathscr{A}, D, D', O))]$ is the log of the moment generating function of the privacy loss. The privacy guarantee of the composite mechanism $\mathscr{A}_{1:k}$ can be computed using that $\alpha_{\mathscr{A}_{1:k}}(\lambda) \leq \sum_{i=1}^k \alpha_{\mathscr{A}_i}(\lambda)$ Abadi et al. [2016].

Gaussian Mechanism. A fundamental concept of all DP sanitization techniques is the *global sensitivity* of a function [Dwork and Roth, 2014].

Definition 3 (Global L_p-sensitivity)**.** *For any function $f : \mathscr{D} \to \mathbb{R}^n$, the L_p-sensitivity of f is $\Delta_p f = \max_{D,D'} ||f(D) - f(D')||_p$, for all D, D' differing in at most one record, where $||\cdot||_p$ denotes the L_p-norm.*

The Gaussian Mechanism [Dwork and Roth, 2014] consists of adding Gaussian noise to the true output of a function. In particular, for any function $f : \mathscr{D} \to \mathbb{R}^n$, the Gaussian mechanism is defined as adding i.i.d Gaussian noise with variance $(\Delta_2 f \cdot \sigma)^2$ and zero mean to each coordinate value of $f(D)$. Recall that the pdf of the Gaussian distribution with mean μ and variance ξ^2 is $\mathsf{pdf}_{\mathscr{G}(\mu,\xi)}(x) = \frac{1}{\sqrt{2\pi}\xi} \exp\left(-\frac{(x-\mu)^2}{2\xi^2}\right)$.

In fact, the Gaussian mechanism draws vector values from a multivariate spherical (or isotropic) Gaussian distribution which is described by random variable $\mathscr{G}(f(D), \Delta_2 f \cdot \sigma \mathbf{I}_n)$, where n is omitted if its unambiguous in the given context.

3 FEDERATED PRUNING

In the standard federated learning scheme (FL-STD, in Section 2), the server sends the latest updated model to a randomly selected set of clients (downstream), and each client sends back its complete model update after local training to the server (upstream) at each round. Knowing that a model has on average millions of parameters (each is a floating point value represented on 32 bits), the network can suffer from large traffic both upstream and downstream.

Our solution, called FL-TOP, aims to reduce the large amount of network traffic by reducing both downstream and upstream traffic. Moreover, a privacy-preserving extension of this scheme, called FL-TOP-DP, is also proposed, which provides Differential Privacy for the whole training data of every client.

In what follows, we first describe the non-private scheme FL-TOP and then the privacy-preserving FL-TOP-DP.

3.1 FL-TOP: FEDERATED PRUNING FOR COMPRESSION

FL-TOP is inspired by the pruning techniques proposed in Han et al. [2016] (see Section 5 for more details), and it aims to reduce the amount of parameters exchanged downstream (from the server to the participating entities) and upstream (from the participating entities to the server). In our scheme, each client updates only a small subset, Top-K, of the model parameters (weights) at each round. Only the K weight values of these Top-K parameters are updated during training, and neither the clients nor the server need to transfer the values of the remaining $n-K$ parameters, where n is the total number of parameters. The set of Top-K parameters do not change over the whole training and are identical for all clients. We experimentally show in Section 4 that, if these K parameters are chosen carefully, the performance penalty is negligible even if $K = 0.005 \cdot n$, that is, 99.5% of the model parameters are pruned. Note that unlike standard pruning techniques, where the set of pruned weights are re-selected after each SGD iteration [Han et al., 2016], our scheme always updates the same K parameters.

These Top-K parameters are selected by the server at the beginning of the protocol. More specifically, the server initializes the model and trains that with some public data that have a similar distribution as the clients' training data. After a few SGD iterations, the server selects the K parameters which values changed the most.

FL-TOP is described in Alg. 1. First, the server uses public data to identify the set $\mathbb{T}$ of the Top-K parameters ($K = |\mathbb{T}|$), before starting federated learning. In particular, starting from a public model $\mathbf{w}_0$, it accumulates the absolute value of gradients per parameter over T_{init} SGD iterations, and selects the K parameters with the largest accumulated gradients. After that, the values/updates[1] of these parameters are the only ones exchanged during the rest of the training between the server and the clients.

At each round, each selected client k uses the K updated weights $\hat{\mathbf{w}}_{t-1}$ received from the server to create a new weight vector $\mathbf{w}^k_{t-1}$ of size n, such that $\mathbf{w}^k_{t-1}$ is composed from the compressed vector $\hat{\mathbf{w}}^k_{t-1}$ of size $K \leq n$ (with coordinates in $\mathbb{T}$) and $n-K$ weights from the initialization vector $\mathbf{w}_0$. $\mathbf{w}_0$ is identical for all participants and can be generated from a shared seed. Note that when $K = |\mathbb{T}| = n$, the scheme is equivalent to FL-STD. The weight vector $\mathbf{w}^k_{t-1}$ is used to train the client's model. However, only the weights in $\mathbb{T}$ are updated while the remaining ones are kept fixed. To do that, the weights not in $\mathbb{T}$ are reinitialized after each SGD iteration to $\mathbf{w}_0$. The server receives only the values from $\mathbf{w}^k_t - \mathbf{w}^k_{t-1}$ at coordinates $\mathbb{T}$, denoted by $\mathscr{C}(\mathbf{w}^k_t - \mathbf{w}^k_{t-1})$ for short, from every client k, and updates the common model $\mathbf{w}_t$ with the average of these compressed updates (in Line 12).

3.2 FL-TOP-DP: DIFFERENTIALLY PRIVATE FEDERATED PRUNING

3.2.1 Privacy Model

We consider an adversary, or a set of colluding adversaries, who can access any update vector sent by the server or any clients at each round of the protocol. A plausible adversary is a participating entity, i.e. a malicious client or server, that wants to infer the training data used by other participants. The adversary is *passive* (i.e., honest-but-curious), that is, it follows the learning protocol faithfully.

Different privacy requirements can be considered depending

[1] weight values for downstream and update/gradients for upstream traffic

on what information the adversary aims to infer. In general, private information can be inferred about:

- any record (user) in any dataset of any client (*record-level privacy*),
- any client/party (*client-level privacy*).

To illustrate the above requirements, suppose that several banks build a common model to predict the creditworthiness of their customers. A bank certainly does not want other banks to learn the financial status of any of their customers (record privacy) and perhaps not even the average income of all their customers (client privacy).

Record-level privacy is a standard requirement used in the privacy literature and is usually weaker than client-level privacy. Indeed, client-level privacy requires to hide any information which is unique to a client including perhaps all its training data.

We aim at developing a solution that provides *client-level privacy and is also bandwidth efficient*. For example, in the scenario of collaborating banks, we aim at protecting any information that is unique to each single bank's training data. The adversary should not be able to learn from the received model or its updates whether any client's data is involved in the federated run (up to ε and δ). We believe that this adversarial model is reasonable in many practical applications when the confidential information spans over multiple samples in the training data of a single client (e.g., the presence of a group a samples, such as people from a certain race). Differential Privacy guarantees plausible deniability not only to any groups of samples of a client but also to any client in the federated run. Therefore, any negative privacy impact on a party (or its training samples) cannot be attributed to their involvement in the protocol run.

3.2.2 Operation

FL-TOP-DP is described in Alg. 3 is very similar to FL-TOP except that each client adds Gaussian noise to its Top-K model updates to guarantee client-level DP, and applies secure aggregation allowing the server to learn only the aggregated (and noisy) model update. More specifically, each client first calculates its compressed model update $\Delta\mathbf{w}_t^k = \mathscr{C}(\mathbf{w}_t^k - \mathbf{w}_{t-1}^k)$ (in Line 25) which is then clipped (in Line 26) to obtain $\Delta\hat{\mathbf{w}}_t^k$ with L_2-norm at most S. After that, random noise $\mathbf{z}_k \sim \mathscr{G}(0, S\sigma\mathbf{I}/\sqrt{\mathbb{K}})$ is added to $\Delta\hat{\mathbf{w}}_t^k$ such that the sum $\sum_{k\in\mathbb{K}}(\Delta\hat{\mathbf{w}}_t^k + \mathbf{z}_k) = \sum_{k\in\mathbb{K}}\Delta\hat{\mathbf{w}}_t^k + \mathscr{G}(0, S\sigma\mathbf{I})$ as the sum of Gaussian random variables also follows Gaussian distribution[2] and then differential privacy is satisfied where ε and δ can be computed using the moments accountant described in Section 2.2. Recall that the Top-K coordinates in $\mathbb{T}$ are selected and distributed by the server, which is honest-but-curious by assumption.

However, as the noise is inversely proportional to $\sqrt{\mathbb{K}}$, $\mathbf{z}_k$ is likely to be small if $|\mathbb{K}|$ is too large. Therefore, the adversary accessing an individual update $\Delta\hat{\mathbf{w}}_t^k + \mathbf{z}_k$ can almost learn a non-noisy update since $\mathbf{z}_k$ is small. Hence, each client uses secure aggregation to encrypt its individual update before sending it to the server. Upon reception, the server sums the encrypted updates as:

$$\begin{aligned}\sum_{k\in\mathbb{K}} \mathbf{c}_t^k &= \sum_{k\in\mathbb{K}} \mathsf{Enc}_{K_k}(\Delta\hat{\mathbf{w}}_t^k + \mathbf{z}_k) = \sum_{k\in\mathbb{K}} \Delta\hat{\mathbf{w}}_t^k + \sum_{k\in\mathbb{K}} \mathbf{z}_k \\ &= \sum_{k\in\mathbb{K}} \Delta\hat{\mathbf{w}}_t^k + \mathscr{G}(0, S\sigma\mathbf{I}) \qquad (1)\end{aligned}$$

where $\mathsf{Enc}_{K_k}(\Delta\hat{\mathbf{w}}_t^k + \mathbf{z}_k) = \Delta\hat{\mathbf{w}}_t^k + \mathbf{z}_k + \mathbf{K}_k \mod p$ and $\sum_k \mathbf{K}_k = 0$ (see Ács and Castelluccia [2011], Bonawitz et al. [2016] for details). Here the modulo is taken element-wise and $p = 2^{\lceil \log_2(\max_k ||\Delta\hat{\mathbf{w}}_t^k + \mathbf{z}_k||_\infty |\mathbb{K}|)\rceil}$. Let $\gamma_t^k = 1/\max\left(1, \frac{||\Delta\mathbf{w}_t^k||_2}{S}\right)$. Then,

$$\begin{aligned}\sum_{k\in\mathbb{K}} \Delta\hat{\mathbf{w}}_t^k &= \sum_{k\in\mathbb{K}} \gamma_t^k \Delta\mathbf{w}_t^k = \sum_{k\in\mathbb{K}} \gamma_t^k \mathscr{C}(\mathbf{w}_t^k - \mathbf{w}_{t-1}^k, \mathbb{T}) \\ &= \mathscr{C}(\sum_{k\in\mathbb{K}} \gamma_t^k (\mathbf{w}_t^k - \mathbf{w}_{t-1}^k), \mathbb{T}) \qquad (2)\end{aligned}$$

where the last equality comes from the linearity of the compression operation. Indeed, recall that each client selects the values of the *same* Top-K coordinates from $\mathbb{T}$. Plugging Eq. (2) into Eq. (1). we get that

$$\sum_{k\in\mathbb{K}} \mathbf{c}_t^k = \mathscr{C}(\sum_{k\in\mathbb{K}} \gamma_t^k (\mathbf{w}_t^k - \mathbf{w}_{t-1}^k), \mathbb{T}) + \mathscr{G}(0, S\sigma\mathbf{I})$$

Privacy analysis: The server can only access the noisy aggregate which is sufficiently perturbed to ensure differential privacy; any client-specific information that could be inferred from the noisy aggregate is tracked and quantified by the moments accountant, described in Section 2.2, as follows.

Let $\eta_0(x|\xi) = \mathsf{pdf}_{\mathscr{G}(0,\xi)}(x)$ and $\eta_1(x|\xi) = (1-C)\mathsf{pdf}_{\mathscr{G}(0,\xi)}(x) + C\mathsf{pdf}_{\mathscr{G}(1,\xi)}(x)$ where C is the sampling probability of a single client in a single round. Let $\alpha(\lambda|C) = \log\max(E_1(\lambda,\xi,C), E_2(\lambda,\xi,C))$ where $E_1(\lambda,\xi,C) = \int_{\mathbb{R}} \eta_0(x|\xi,C) \cdot \left(\frac{\eta_0(x|\xi,C)}{\eta_1(x|\xi,C)}\right)^\lambda dx$ and $E_2(\lambda,\xi,C) = \int_{\mathbb{R}} \eta_1(x|\xi,C) \cdot \left(\frac{\eta_1(x|\xi,C)}{\eta_0(x|\xi,C)}\right)^\lambda dx$.

Theorem 1 (Privacy of FL-TOP-DP)**.** *FL-TOP-DP is* $(\min_\lambda(T_{\mathsf{cl}} \cdot \alpha(\lambda|C) - \log\delta)/\lambda, \delta)$*-DP.*

Given a fixed value of δ, ε is computed numerically as in Abadi et al. [2016], Mironov et al. [2019].

[2] More precisely, $\sum_i \mathscr{G}(\nu_i, \xi_i) = \mathscr{G}(\sum_i \nu_i, \sqrt{\sum_i \xi_i^2})$

Algorithm 1: FL-TOP: Federated Learning

Server:
  Initialize common model w_0
  Select set $\mathbb{T}$ of Top-K updated weights' coordinates via public dataset
  for $t = 1$ **to** T_{cl} **do**
    Select $\mathbb{K}$ clients uniformly at random
    for *each client k in* $\mathbb{K}$ **do**
      $\mathbf{c}_t^k = \textbf{Client}_k(\mathscr{C}(\mathbf{w}_{t-1}, \mathbb{T}))$
    end
    $\mathbf{w}_t = \mathbf{w}_0$
    $j = 1$
    for *each coordinate i in* $\mathbb{T}$ **do**
      $\mathbf{w}_t[i] = \mathbf{w}_{t-1}[i] + \sum_k \frac{\mathbf{c}_t^k[j]}{|\mathbb{K}|}$
      $j = j+1$
    end
  end
  Output: Global model $\mathbf{w}_t$

Client$_k(\hat{\mathbf{w}}_{t-1}^k)$**:**
  $\mathbf{w}_{t-1}^k = \mathbf{w}_0$
  $j = 1$
  for *each coordinate i in* $\mathbb{T}$ **do**
    $\mathbf{w}_{t-1}^k[i] = \hat{\mathbf{w}}_{t-1}^k[j]$
    $j = j+1$
  end
  $\mathbf{w}_t^k = \textbf{Top}_\mathbf{k}\textbf{SGD}(D_k, \mathbf{w}_{t-1}^k, \mathbf{w}_0, T_{\mathsf{gd}}, \mathbb{T})$
  Output: Model update $\mathscr{C}(\mathbf{w}_t^k - \mathbf{w}_{t-1}^k, \mathbb{T})$

Algorithm 2: **Top$_\mathbf{k}$**-Stochastic Gradient Descent

Input: D : training data, T_{gd} : local epochs, $\mathbf{w}$: weights, $\mathbf{w}_0$: first weights' initialization, $\mathbb{T}$: set of Top-K values coordinates .

for $t = 1$ **to** T_{gd} **do**
  Select batch $\mathbb{B}$ from D randomly
  $\mathbf{u} = -\eta \nabla f(\mathbb{B}; \mathbf{w})$
  for *each coordinate i in* $\mathbb{T}$ **do**
    $\mathbf{w}[i] = \mathbf{w}[i] + \mathbf{u}[i]$
  end
end
Output: Model $\mathbf{w}$

3.2.3 Remarks

The magnitude of the added Gaussian noise is proportional to the clipping threshold S, which is in turn calibrated to the norm of the model update. However, the norm of the model update increases if the model size increases [Zhu et al., 2020], and hence S should be chosen sufficiently large to guarantee fast convergence with large accuracy. On the other hand, too large S also increases the perturbation error caused by the added noise.

FL-TOP aims to diminish this perturbation error by reducing S via compression which also increases the L_2-norm of the compressed update vector. This is illustrated in Figure 1, which shows that the norm of the Top-K coordinates with FL-TOP tend to be larger than with FL-STD (i.e., when all coordinates get updated not only the Top-K). Therefore, besides decreasing the magnitude of the added noise, FL-TOP also decreases the relative error on the retained parameters. These together decrease the perturbation error caused by the added noise.

Notice that there exist other alternatives to identify the Top-K coordinates in a privacy-preserving manner than using a public dataset. For example, every client can select the Top-K parameters with the largest magnitude during the first rounds locally, and send them to the server for aggregation. More specifically, each client creates a parameter vector with size n, where the Top-K coordinates are set to 1 while the rest are kept 0. Then, these binary vectors are noised and aggregated by the server like in Section 3.2.2. In the rest of the training, all participants exchange only the updates and weights of the these Top-K parameters like in FL-TOP. However, aside from consuming more privacy budget, this approach also has lower accuracy than our proposal according to our tests. Moreover, it has larger communication cost in the initialization phase when the Top-K parameters are identified and the whole binarized parameter vector is sent for aggregation.

4 EXPERIMENTAL RESULTS

The goal of this section is to evaluate the performance of our proposed schemes FL-TOP and FL-TOP-DP on a benchmark dataset and a realistic in-hospital mortality prediction scenario. We aim at evaluating their performance with different levels of compression and comparing them with the performance of the following learning protocols[3]:

- FL-STD: The Standard Federated Learning scheme as described in Section 2.1 (see Alg. 4).
- FL-BASIC: A Federated Learning scheme that updates a random subset of parameters instead of the Top-K parameters at each SGD iteration. This subset is re-selected at the beginning of each new round. The $n-k$ non-selected parameters are still reinitialized after each SGD update as in FL-TOP.
- FL-CS: A Federated Learning scheme that uses Compressive sensing (CS) to compress model updates from Kerkouche et al. [2020]. See Section 5 for more details.

Note that all compression operators in the baselines are linear (just like FL-TOP-DP), and hence they can also be used with secure aggregation. Similarly to FL-TOP-DP, the

[3]More baselines are considered but due to the lack of space, we have decided to present only those which return the best results. All other results can be found in the appendix(Section D).

Algorithm 3: FL-TOP-DP: Federated Learning

Server:
  Initialize common model w_0
  Select set $\mathbb{T}$ of Top-K updated weights' coordinates via public dataset
  for $t = 1$ **to** T_{cl} **do**
    Select $\mathbb{K}$ clients uniformly at random
    for *each client k in* $\mathbb{K}$ **do**
      $\mathbf{c}_t^k = \textbf{Client}_k(\mathscr{C}(\mathbf{w}_{t-1}, \mathbb{T}))$
    end
    $\mathbf{w}_t = \mathbf{w}_0$
    $j = 1$
    for *each coordinate i in* $\mathbb{T}$ **do**
      $\mathbf{w}_t[i] = \mathbf{w}_{t-1}[i] + \sum_k \frac{\mathbf{c}_t^k[j]}{|\mathbb{K}|}$
      $j = j+1$
    end
  end
  Output: Global model $\mathbf{w}_t$

Client$_k(\hat{\mathbf{w}}_{t-1}^k)$**:**
  $\mathbf{w}_{t-1}^k = \mathbf{w}_0$
  $j = 1$
  for *each coordinate i in* $\mathbb{T}$ **do**
    $\mathbf{w}_{t-1}^k[i] = \hat{\mathbf{w}}_{t-1}^k[j]$
    $j = j+1$
  end
  $\mathbf{w}_t^k = \textbf{Top}_{\mathbf{k}}\textbf{SGD}(D_k, \mathbf{w}_{t-1}^k, \mathbf{w}_0, T_{\mathsf{gd}}, \mathbb{T})$
  $\Delta\mathbf{w}_t^k = \mathscr{C}(\mathbf{w}_t^k - \mathbf{w}_{t-1}^k, \mathbb{T})$
  $\Delta\hat{\mathbf{w}}_t^k = \Delta\mathbf{w}_t^k / \max\left(1, \frac{||\Delta\mathbf{w}_t^k||_2}{S}\right)$
  Output: $\mathsf{Enc}_{K_k}(\mathscr{G}(\Delta\hat{\mathbf{w}}_t^k, S\mathbf{I}\sigma/\sqrt{|K|}))$

private extensions (i.e., FL-STD-DP, FL-BASIC-DP and FL-CS-DP) also clip and then noise the compressed updates.

We evaluate the above learning algorithms on the well-known Fashion-MNIST dataset [Xiao et al., 2017] and on the Premier Healthcare Database, which is a real-world medical dataset of 1.2 millions of US hospital patients [4]. More details can be found in Appendix A.1 and Appendix B.1.

Recall that the Top-K weights are selected before starting the federated learning process using public data. For Fashion-MNIST, we randomly select a batch with size 10 from MNIST dataset [LeCun and Cortes, 2010] described in Appendix B.2. For the medical dataset, we did not find any public dataset with the same features as ours, and for this reason, we selected randomly from the dataset a batch of 356 patients[5]. This set is used only by the server and never by any client. Afterwards, the server performs T_{init} SGD iterations starting from the model parameters $\mathbf{w}_0$ on the same batch to identify the Top-K weights. We experimentally show later that even these small batches are enough for the server to find a good set of Top-K weights.

In order to select the clipping threshold S, the server executes a single training round locally, which is composed of T_{gd} SGD iterations starting from the model parameters $\mathbf{w}_0$, using the batch from the public data. The clipping threshold S is set to the L_2-norm of the Top-K weight update obtained for this single training round. For FL-BASIC-DP, the same steps are repeated for 100 times, where a new random set of trainable weights with size K are selected each time, which yields 100 L_2-norm values. S is set to the median of these L_2-norm values. We think that this approach is more fair, because the set of trainable weights is re-selected at each round in FL-BASIC-DP. The computed values of S can be found in Table 6 and Table 7 for Fashion-MNIST and Medical dataset, respectively. More information about the model architectures and the hyper-parameter selection can be found in Appendix A.

4.1 RESULTS

Figure 1 displays the distribution of the Top-K updated weights for FL-TOP and FL-STD at the end of the training. We select the weights when each scheme reached the best accuracy over 200 and best balanced accuracy[6] over 100 rounds for fashion-MNIST and the medical dataset, respectively. We choose the smallest compression ratio r that leads to the best accuracy for the FL-TOP-DP scheme. Table 1 shows that FL-TOP-DP reaches the best accuracy, 0.81, when $r = 0.5\%$ on fashion-MNIST and reaches the best accuracy, 0.69, when $r = 0.1\%$ on the medical dataset. Both figures validate the intuition that by constraining the model to update only a small set K of the total weights, these Top-K become more important and reach larger values. This result is important when differential privacy is used as it leads to larger value-to-noise level and therefore better performance.

Table 1 represents the best accuracy over 200 rounds for each scheme on the Fashion-MNIST dataset. *Round* corresponds to the round when the best accuracy is reached and *Cost* is the average bandwidth consumption calculated as: $r \times n \times 32 \times Round \times C$, where 32 is the number of bits necessary to represent a float value, n is the uncompressed model size, $r = \frac{|\mathbb{T}|}{n}$, $|\mathbb{T}|$ is the compressed model size, C is the sampling probability of a client, and *Round* is the round when we get the the best accuracy.

Table 2 represents the best balanced accuracy over 100 rounds for each scheme on the Medical dataset. *AUROC* (area under the receiver operating characteristic curve - see Appendix A.4) corresponds to the *AUROC* value when the best balanced accuracy is reached.

[4] https://www.premierinc.com/newsroom/education/premier-healthcare-database-whitepaper

[5] Reduced to 24 patients when we train via downsampling with 12 patients for each class

[6] See Appendix A.4 for more details.

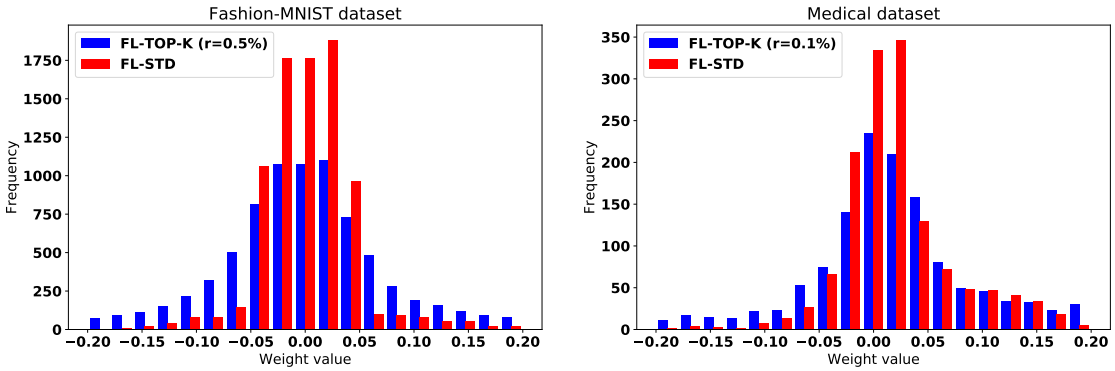

Figure 1: Distributions of the Top-K weight values (after convergence) for both FL-TOP and FL-STD schemes with the Fashion-MNIST dataset (left) and the medical dataset (right).

r	Algorithms	*Performance*				
		Accuracy	*Round*	*Downstream Cost* (Kilobyte)	*Upstream Cost* (Kilobyte)	ε
0.5%	FL-BASIC	0.65	193	21402.03	107	N/A
	FL-CS	0.57	185	20514.9	102.56	N/A
	FL-TOP	**0.82**	200	**110.88**	**110.88**	N/A
	FL-BASIC-DP	0.59	200	22178.27	110.88	1
	FL-CS-DP	0.53	200	22178.27	110.88	1
	FL-TOP-DP	**0.81**	200	**110.88**	**110.88**	**1**
5%	FL-BASIC	0.78	196	21734.70	1086.73	N/A
	FL-CS	0.82	200	22178.27	1108.91	N/A
	FL-TOP	**0.84**	200	**1108.91**	**1108.91**	N/A
	FL-BASIC-DP	0.76	195	21623.81	1081.18	0.99
	FL-CS-DP	0.78	160	17742.61	887.13	0.94
	FL-TOP-DP	**0.81**	152	**842.77**	**842.77**	**0.92**
10%	FL-BASIC	0.81	196	21734.70	2173.47	N/A
	FL-CS	0.85	182	20182.22	2018.22	N/A
	FL-TOP	**0.85**	199	**2206.74**	**2206.74**	N/A
	FL-BASIC-DP	0.79	189	20958.46	2095.85	0.98
	FL-CS-DP	0.72	167	18518.85	1851.89	0.95
	FL-TOP-DP	**0.80**	157	**1740.99**	**1740.99**	**0.93**
100%	FL-STD	0.86	200	22178.27	22178.27	N/A
	FL-STD-DP	0.56	60	6653.48	6653.48	0.76

Table 1: Summary of results on Fashion-MNIST dataset.

r	Algorithms	*Performance*					
		Bal_Acc	*AUROC*	*Round*	*Downstream Cost* (Kilobyte)	*Upstream Cost* (Kilobyte)	ε
0.1%	FL-BASIC	0.51	0.51	99	11829.42	11.82	N/A
	FL-CS	0.53	0.55	100	11948.91	11.94	N/A
	FL-TOP	**0.69**	**0.76**	68	**8.12**	**8.12**	N/A
	FL-BASIC-DP	0.50	0.49	100	11948.91	11.94	1
	FL-CS-DP	0.51	0.51	99	11829.42	11.82	1
	FL-TOP-DP	**0.69**	**0.76**	85	**10.15**	**10.15**	**0.97**
5%	FL-BASIC	0.72	0.80	100	11948.91	597.45	N/A
	FL-CS	0.73	0.81	98	11709.93	585.5	N/A
	FL-TOP	**0.72**	**0.80**	95	**567.57**	**567.57**	N/A
	FL-BASIC-DP	0.69	0.76	100	11948.91	597.45	1
	FL-CS-DP	0.69	0.76	100	11948.91	597.45	1
	FL-TOP-DP	**0.68**	**0.75**	23	**137.41**	**137.41**	**0.79**
10%	FL-BASIC	0.74	0.81	100	11948.91	1194.89	N/A
	FL-CS	0.74	0.82	100	11948.91	1194.89	N/A
	FL-TOP	**0.74**	**0.82**	90	**1075.40**	**1075.40**	N/A
	FL-BASIC-DP	0.69	0.76	99	11829.42	1182.94	1
	FL-CS-DP	0.69	0.76	96	11470.95	1147.09	0.99
	FL-TOP-DP	**0.68**	**0.74**	23	**274.82**	**274.82**	**0.79**
100%	FL-STD	0.74	0.82	99	11829.42	11829.42	N/A
	FL-STD-DP	0.66	0.72	62	7408.32	7408.32	0.91

Table 2: Summary of results on Medical dataset.

These tables show that the proposed non-private scheme FL-TOP has similar accuracy than the standard scheme FL-STD but reduces the bandwidth cost significantly. For example, with the Fashion-MNIST dataset, the FL-TOP accuracy reaches 0.85 when the compression ratio $r = 10\%$. In comparison, the standard FL-STD scheme reaches an accuracy of 0.86% but consumes 10 times more bandwidth. Furthermore, although FL-CS reaches the same accuracy than FL-TOP and consumes slightly less bandwidth upstream (9% less), its required downstream bandwidth is about 10 times larger (See Table 1 for more details). The results on the medical dataset are quite similar. In fact, FL-TOP achieves its best balanced accuracy (0.74) and AUROC (0.82) when $r = 10\%$ while the FL-STD scheme obtains similar performance but required about 11 times more upsteam and downstream bandwidth cost. FL-CS achieves similarly accuracy at $r = 10\%$ as FL-TOP but its downstream required bandwidth is about 11 times larger (see Table 2 for more details).

The results also show that not only our privacy-preserving solution FL-TOP-DP provides strong privacy guarantee (with ε values smaller than 1) but that it outperforms the other schemes in term of accuracy and bandwidth, for both datasets. For example, with Fashion-MNIST, our scheme achieves an accuracy of 0.81 when $r = 0.5\%$ while the baseline scheme, FL-BASIC-DP, achieves an accuracy of 0.79 when $r = 10\%$ and requires 189 times more downstream bandwidth and 18 times more upstream bandwidth. With the medical dataset, FL-TOP-DP reaches the best balanced accuracy 0.69 and best AUROC 0.76 for a compression ratio of $r = 0.1\%$ while FL-BASIC-DP and FL-CS-DP achieves the same performance at $r = 5\%$. Note that FL-STD-DP performs very poorly as noise has to be added to the all weights of the model and the sensitivity is large (see Table 2).

5 RELATED WORK

Privacy of Federated Learning: The concept of Client-based Differential Privacy has been introduced in McMahan et al. [2018] and Geyer et al. [2017], where the goal is to hide any information that is specific to a single client's training data. These algorithms noise the contribution of a single client instead of a single record in the client's dataset. The noise is added by the server, hence, unlike our solution,

these works assume that the server is trusted.

Recently, Liu et al. [2020] also proposed to add noise only to the update of the Top-K model parameters a la local-DP. In local-DP, each client adds larger noise that what is necessary to guarantee DP for the *aggregated* model update without using secure aggregation. Therefore, the common model is less accurate than with our scheme. In addition, Liu et al. [2020] uses two epsilon budgets; one for selecting Top-K parameters per client, and the second for perturbing these selected Top-K parameters. By contrast, we select the Top-K parameters via public data without sacrificing any privacy budget. Finally, their solution is also less bandwidth efficient than ours: as the Top-K parameters differ for each client and at each round, the client cannot send only the Top-K parameters values because the server will not be able to identify which value corresponds to which Top-K parameter. For this reason, the client has to send a sparse vector with only Top-K perturbed values and all remaining parameters set to 0. Therefore, the quantization of the non-Top-K parameters is performed only during the upstream (from client to server) without compressing any downstream traffic. As opposed to this, in our solution, only the weights/updates of the Top-K parameters are transferred downstream/upstream.

Recently, Kerkouche et al. [2020] proposed to use Compressive sensing (CS) in the context of federated learning in order to compress model updates meanwhile providing client-level DP. Assuming that the model update is already sparse in the time domain, the noise is added to its largest Fourier coefficients in a distributed manner, and the noisy aggregate is reconstructed with standard optimization techniques. Likewise our solution, this work also uses secure aggregation by exploiting the linearity of CS. However, the reconstruction process can be slow for large models, and therefore our solution is more scalable. Moreover, it can only compress the upstream traffic.

Bandwidth Optimization in Federated Learning: Different quantization methods have been proposed to save the bandwidth and reduce the communication costs in federated learning. They can be divided into two main groups: unbiased and biased methods. The unbiased approximation techniques use probabilistic quantization schemes to compress the stochastic gradient and attempt to approximate the true gradient value as much as possible [Alistarh et al., 2016, Wen and al., 2017, Wang et al., 2018, Konecný et al., 2016]. However, biased approximations of the stochastic gradient can still guarantee convergence both in theory and practice [Bernstein et al., 2018, Lin et al., 2018, Seide et al., 2014]. SignSGD Bernstein et al. [2018] a quantization protocol allows to compress during downstream and upstream traffic but requires the use of all the clients at each round which is not realistic in the context of federated settings because each client is available only during few rounds Kairouz et al. [2019].

A different line of works exploit the sparsity of model updates to compress them. Amiri and Gündüz [2019a,b] proposed to use a compressive sensing for federated learning in order to compress model updates without privacy guarantees. However, they assume that all clients participate in each round (as they maintain an error accumulation vector at each client due to the compression scheme), but as discussed in Kairouz et al. [2019] this assumption is not always realistic. Sketching was adapted to federated learning for the purpose of compressing model updates in Ivkin et al. [2019] and Rothchild et al. [2020]. The authors proposed to use Count-Sketch from Charikar et al. [2002] to retrieve the largest weights in the update vector on the server side. However, it is unclear how these works can be extended with privacy guarantees. Moreover, unlike our technique, they do not compress downstream traffic.

Constraining the weights to have a specific distribution has already been studied. In Han et al. [2016], for example, the authors use pruning techniques to create a sparse model at the end of the training. After each SGD iteration, the authors zero-out all the weights with an absolute value smaller than a threshold. Iterating the process leads to a sparse model with only some absolute weight values larger than 0. Similarly, Courbariaux et al. [2016] aim to create a model with binary weights such that at the end of the training all the weights are close to 1 or -1. After each SGD update, the authors take the sign of the weights before the next update. After some iterations, the weight values become close to the interval limits -1 and 1.

In Frankle and Carbin [2018], a new hypothesis claims that there exists a sub-network which, if trained separately, can achieve similar performance as the complete network model which contains that. To find such a sub-network, one has to follow a simple iterative procedure: train the complete network, prune the smallest weights, and then reinitialize the remaining weights to their original values. These steps are repeated iteratively. This approach was extended to federated learning in Li et al. [2020].

6 CONCLUSION

This paper presents a novel privacy-preserving federated learning scheme that reduces bandwidth, latency and therefore power consumption. The proposed scheme is based on Differential Privacy and therefore provides theoretical privacy guarantees. Furthermore, it optimizes the model accuracy by constraining the model learning phase on few selected weights. We show experimentally, using a public dataset called Fashion-MNIST and a real world medical dataset of 1.2 millions of US hospital patients, that it reduces the upstream and downstream bandwidth by up to 99.9% compared to standard federated learning, making it practical for constrained and mobile devices.

References

Martín Abadi, , et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *ACM CCS*, 2016.

Gergely Ács and Claude Castelluccia. I have a dream! (differentially private smart metering). In *IH*, 2011.

Josephine Akosa. Predictive accuracy: a misleading performance measure for highly imbalanced data. In *Proceedings of the SAS Global Forum*, pages 2–5, 2017.

Dan Alistarh, Jerry Li, Ryota Tomioka, and Milan Vojnovic. QSGD: randomized quantization for communication-optimal stochastic gradient descent. 2016.

Mohammad Mohammadi Amiri and Deniz Gündüz. Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air. 2019a.

Mohammad Mohammadi Amiri and Deniz Gündüz. Federated learning over wireless fading channels. 2019b.

Anand Avati, Kenneth Jung, Stephanie Harman, Lance Downing, Andrew Ng, and Nigam H. Shah. Improving palliative care with deep learning. *BMC Medical Informatics and Decision Making*, 2018.

Mohamed Bekkar, Hassiba Djema, and T.A. Alitouche. Evaluation measures for models assessment over imbalanced data sets. *Journal of Information Engineering and Applications*, 3:27–38, 01 2013.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Anima Anandkumar. signsgd: compressed optimisation for non-convex problems. 2018.

Keith Bonawitz et al. Practical secure aggregation for federated learning on user-held data. 2016.

Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M Buhmann. The balanced accuracy and its posterior distribution. In *20th International Conference on Pattern Recognition*. IEEE, 2010.

Moses Charikar, Kevin Chen, and Martin Farach-Colton. Finding frequent items in data streams. In *International Colloquium on Automata, Languages, and Programming*, pages 693–703. Springer, 2002.

François Chollet et al. Keras. `https://keras.io`, 2015a.

François Chollet et al. Keras datasets. `https://keras.io/datasets/`, 2015b.

Matthieu Courbariaux, Itay Hubara, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks: Training deep neural networks with weights and activations constrained to +1 or -1, 2016.

Marta TERRON CUADRADO. Icd-9-cm: International classification of diseases, ninth revision, clinical modification. `https://ec.europa.eu/cefdigital/wiki/display/EHSEMANTIC/ICD-9-CM%3A+International+Classification+of+Diseases%2C+Ninth+Revision%2C+Clinical+Modification`, 2019.

Cynthia Dwork and Aaron Roth. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4), 2014.

Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. RAPPOR: randomized aggregatable privacy-preserving ordinal response. In Gail-Joon Ahn, Moti Yung, and Ninghui Li, editors, *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, Scottsdale, AZ, USA, November 3-7, 2014*, pages 1054–1067. ACM, 2014. doi: 10.1145/2660267.2660348. URL `https://doi.org/10.1145/2660267.2660348`.

A. Fejza, P. Genevès, N. Layaïda, and J. Bosson. Scalable and interpretable predictive models for electronic health records. In *DSAA*, 2018.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Training pruned neural networks. 2018.

Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients – how easy is it to break privacy in federated learning?, 2020.

Robin C. Geyer, Tassilo Klein, and Moin Nabi. Differentially private federated learning: A client level perspective. 2017.

Song Han, Jeff Pool, Sharan Narang, Huizi Mao, Enhao Gong, Shijian Tang, Erich Elsen, Peter Vajda, Manohar Paluri, John Tran, Bryan Catanzaro, and William J. Dally. Dsd: Dense-sparse-dense training for deep neural networks, 2016.

Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 2009.

Nikita Ivkin, Daniel Rothchild, Enayat Ullah, Ion Stoica, Raman Arora, et al. Communication-efficient distributed sgd with sketching. In *Advances in Neural Information Processing Systems*, pages 13144–13154, 2019.

Peter Kairouz et al. Advances and open problems in federated learning. 2019.

Raouf Kerkouche, Gergely Ács, Claude Castelluccia, and Pierre Genevès. Compression boosts differentially private federated learning, 2020. To appear in EuroS&P 2021.

Jakub Konecný, H. Brendan McMahan, Felix X. Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. 2016.

Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010. URL `http://yann.lecun.com/exdb/mnist/`.

Ang Li, Jingwei Sun, Binghui Wang, Lin Duan, Sicheng Li, Yiran Chen, and Hai Li. Lotteryfl: Personalized and communication-efficient federated learning with lottery ticket hypothesis on non-iid datasets, 2020.

Yujun Lin, Song Han, Huizi Mao, Yu Wang, and Bill Dally. Deep gradient compression: Reducing the communication bandwidth for distributed training. In *ICLR*, 2018.

Ruixuan Liu, Yang Cao, Masatoshi Yoshikawa, and Hong Chen. Fedsel: Federated sgd under local differential privacy with top-k dimension selection. *Lecture Notes in Computer Science*, 2020.

Rupa Makadia and Patrick B. Ryan. Transforming the premier perspective® hospital database into the observational medical outcomes partnership (omop) common data model. In *EGEMS*, 2014.

Margaret Mcdonald, Timothy Peng, Sridevi Sridharan, Janice Foust, Polina Kogan, Liliana Pezzin, and Penny Feldman. Automating the medication regimen complexity index. *Journal of the American Medical Informatics Association : JAMIA*, 2012.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*, 2016.

H. Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. In *International Conference on Learning Representations*, 2018.

Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Inference attacks against collaborative learning. 2018.

Ilya Mironov, Kunal Talwar, and Li Zhang. Rényi differential privacy of the sampled gaussian mechanism. 2019.

Ajinkya More. Survey of resampling techniques for improving classification performance in unbalanced datasets. 2016.

Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *IEEE Symposium on Security and Privacy*, 2019.

Travis E Oliphant. *A guide to NumPy*, volume 1. Trelgol Publishing USA, 2006.

Alvin Rajkomar and al. Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, 2018.

Daniel Rothchild, Ashwinee Panda, Enayat Ullah, Nikita Ivkin, Ion Stoica, Vladimir Braverman, Joseph Gonzalez, and Raman Arora. Fetchsgd: Communication-efficient federated learning with sketching, 2020.

Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns. In *INTERSPEECH*, 2014.

Reza Shokri and Vitaly Shmatikov. Privacy-preserving deep learning. In *CCS*, 2015.

Hongyi Wang et al. Atomo: Communication-efficient learning via atomic sparsification. In *NeurIPS*, 2018.

Wei Wen and al. Terngrad: Ternary gradients to reduce communication in distributed deep learning. 2017.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. 2017.

Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. idlg: Improved deep leakage from gradients. 2020.

Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *NeurIPS 2019*, 2019.

Yuqing Zhu, Xiang Yu, Yi-Hsuan Tsai, Francesco Pittaluga, Masoud Faraki, Manmohan chandraker, and Yu-Xiang Wang. Voting-based approaches for differentially private federated learning, 2020.

A MEDICAL DATA: DATA PRE-PROCESSING & EXPERIMENTAL SETUP DETAILS

This section describes our medical dataset and the experimental setting which is used to evaluate the accuracy and the privacy of our proposals.

A.1 MEDICAL DATASET

A.1.1 The In-hospital Mortality Prediction Scenario

The ability to accurately predict the risks in the patient's perspectives of evolution is a crucial prerequisite in order to adapt the care that certain patients receive [Fejza et al., 2018].

We consider the scenario where several hospitals are collaborating to train models for in-hospital mortality prediction using our Federated Learning schemes. This well-studied real-world problem consists in trying to precisely identify the patients who are at risk of dying from complications during their hospital stay [Avati et al., 2018, Rajkomar and al., 2018, Fejza et al., 2018]. As commonly found in the literature [Fejza et al., 2018], for such predictions, we focus on hospital admissions of adults hospitalized for at least 3 days, excluding elective admissions.

A.1.2 The Premier Healthcare Database

We used EHR data from the Premier healthcare database[7] which is one of the largest clinical databases in the United States, collecting information from millions of patients over a period of 12 months from 415 hospitals in the USA [Fejza et al., 2018]. These hospitals are supposedly representative of the United States hospital experience [Fejza et al., 2018]. Each hospital in the database provides discharge files that are dated records of all billable items (including therapeutic and diagnostic procedures, medication, and laboratory usage) which are all linked to a given patient's admission [Fejza et al., 2018, Makadia and Ryan, 2014].

The initial snapshot of the database used in our work (before pre-processing step) comprises the EHR data of 1,271,733 hospital admissions. Electronic Health Record (EHR) is a digital version of a patient's paper chart readily available in hospitals. For developing supervised learning and specifically deep learning models, we focus on a specific set of features from EHR data. The features of interest that capture the patients information are summarized in Table 3. There is a total of 24,428 features per patient, mainly due to the variety of drugs possibly served. As in Avati et al. [2018], we also removed all the features which appear on less than 100 patients' records, hence, the number of features was reduced to 7,280 features.

The Medication regimen complexity index (MRCI) [Mcdonald et al., 2012] is an aggregate score computed from a total of 65 items, whose purpose is to indicate the complexity of the patient's situation. The minimum MRCI score for a patient is 1.5, which represents a single tablet or capsule taken once a day as needed (single medication). However the maximum is not defined since the number of medications increases the score [Mcdonald et al., 2012]. In our case, after statistical analysis of our dataset, we consider the MRCI score as ranging from 2 to 60.

Most real datasets like ours are generally imbalanced with a skewed distribution between the classes. In our case, the positive cases (patients who die during their hospital stay) represent only 3% of all patients. Table 4 gives more details about this distribution after the pre-processing step which is discussed in A.2. To deal with this well-known problem, we have decided to use downsampling technique [More, 2016, He and Garcia, 2009], a standard solution used for this purpose, as used in Kerkouche et al. [2020].

A.2 PREPROCESSING

1. **Features normalization**: we extract from the dataset the values of each feature represented in Table 3. For gender, we use one-hot encoding: Male, Female and Unknown. Similarly, for admission type we use 4 features: Emergency, Urgent, Trauma Center, and Unknown [8]. For drugs, we extract 24,419 features which correspond to the different drugs (name and dosage). A given patient receives only a few of the possible drugs served, resulting in a very sparse patient's record. We use a MinMax normalization for age and MRCI in order to rescale the values of these features between 0 and 1 (using MinMaxScaler class of scikit-learn[9]). The labels that we consider are boolean: true means that the patient died during his hospital stay while false means she survived.
2. **Patients filtering**: We consider patient and drug information of the first day at the hospital so that we can make predictions 24 hours after admission (as commonly found in the literature [Rajkomar and al., 2018, Fejza et al., 2018]). We filter out the pregnant and new-born patients because the medication types and admission services are not the same for theses two categories of patients. Our model prediction is built without patients' historical medical data. This has the

[7] https://www.premierinc.com/newsroom/education/premier-healthcare-database-whitepaper

[8] https://www.resdac.org/cms-data/variables/claim-inpatient-admission-type-code-ffs

[9] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html

advantage to require minimum patient's information and to work for new patients.

3. **Hospitals filtering**: The dataset contains 415 hospitals for a total size of 1,271,733 records. We split randomly the dataset into disjoint training and testing data (80% and 20% respectively). The final dataset for testing contains 254,347 patients, with 7,882 deceased patients and 246,465 non-deceased patients (see Table 4).

   Using Client-Level differential privacy requires to add more noise than Record-Level differential privacy, because the privacy purposes are not the same as detailled in Section 2. To reduce the noise (when ε is fixed) and then improve the utility, we have to reduce the number of iterations or to reduce the sampling probability which are the parameters used to compute ε. We therefore have two options to reduce the sampling probability:

   - Reducing the number of clients selected at each round $|\mathbb{K}|$. However this option also decreases the amount of data, and hence have a negative impact on the utility. We therefore preferred to use the next option.
   - Increasing the total number of clients N: we created more hospitals by splitting randomly the training data over 5010 "virtual" hospitals. We also, took care to have at least one in-hospital dead patient per hospital. Each hospital contains 203 patients. 356 patients are used as public dataset to define the Top-K updated weights. We created 5010 hospitals in order to have approximately the same number of patients per hospital, each of them with some in-hospital dead patients. In practise, Client-Level differential privacy is more adapted to an environment with a large set of clients as explained in McMahan et al. [2018], Geyer et al. [2017].

A.3 IMBALANCED DATA

The dataset of each hospital is imbalanced because the proportion of patients that leave the hospital alive is, fortunately, much larger than in-hospital dead patients. To deal with this well-known problem, we have decided to use downsampling technique [More, 2016, He and Garcia, 2009], a standard solution used for this purpose. [10]

A.4 PERFORMANCE METRICS

We use the following metrics:

- *Balanced accuracy* [Brodersen et al., 2010, Bekkar et al., 2013] is computed as $1/2 \cdot (\frac{TP}{P} + \frac{TN}{N}) = \frac{TPR+TNR}{2}$ and is mainly used with imbalanced data. *True Positive Rate* (*TPR*) and *True Negative Rate* (*TNR*): $TPR = \frac{TP}{P}$ and $TNR = \frac{TN}{N}$, where P and N are the number of positive and negative instances, respectively, and TP and TN are the number of true positive and true negative instances. We note that traditional ("non-balanced") accuracy metrics such as $\frac{TP+TN}{P+N}$ can be misleading for very imbalanced data Akosa [2017]: in our dataset, the minority class has only 3% of all the training samples (see Table 4), which means that a biased (and totally useless) model always predicting the majority class would have a (non-balanced) accuracy of 97%.

- The *area under the ROC curve* (*AUROC*) is also a frequently used accuracy metric. The ROC curve is calculated by varying the prediction threshold from 1 to 0, when *TPR* and *FPR* are calculated at each threshold. The area under this curve is then used to measure the quality of the predictions. A random guess has an *AUROC* value of 0.5, whereas a perfect prediction has the largest *AUROC* value of 1.

A.5 EVALUATION METHOD.

First, we split randomly the dataset of each hospital into disjoint training and testing data (80% and 20% respectively). An entire federated run is executed with this split, and all the metrics are evaluated in every round on the union of all clients' testing data. All metric values of the round with the best balanced metric are recorded.

A.5.1 Model architecture

As in Avati et al. [2018], Kerkouche et al. [2020], we use a fully connected neural network model with the following architecture: two hidden layers of 200 units, which use a Relu activation function followed by an output layer of 1 unit with sigmoid activation function and a binary cross entropy loss function. This results in 1,496,601 parameters in total. We tune η from 0.01 to 0.5 with an increment value of 0.005. As in Kerkouche et al. [2020], we fix the momentum parameter ρ to 0.9 and the global learning rate η_G to 1.0. The number of chunks is set to $P = 100$ (refers to Kerkouche et al. [2020] for details). The hyperparameters used by each of the considered schemes are summarized in Table 5.

[10] We have also tested weighted loss function and oversampling techniques. But, we noticed experimentally that downsampling technique outperforms the other techniques for all the schemes.

B FASHION-MNIST DATA: DATA PRE-PROCESSING & EXPERIMENTAL SETUP DETAILS

B.1 DATA DESCRIPTION

Fashion-MNIST database of fashion articles consists of 60,000 28x28 grayscale images of 10 fashion categories, along with a test set of 10,000 images Xiao et al. [2017] Chollet et al. [2015b].

B.2 PUBLIC DATA DESCRIPTION

The MNIST database of handwritten digits. It consists of 28 x 28 grayscale images of digit items and has 10 output classes. The training set contains 60,000 data samples while the test/validation set has 10,000 samples LeCun and Cortes [2010] Chollet et al. [2015b].

B.3 PREPROCESSING

The pixel of each image is an unsigned integer in the range between 0 and 255. We rescale them to the range [0,1] instead.

B.4 MODEL ARCHITECTURE

For Fashion-MNIST, we use a model McMahan et al. [2016], Kerkouche et al. [2020] with the following architecture: a convolutional neural network (CNN) with two 5x5 convolution layers (the first with 32 filters, the second with 64, each followed with 2x2 max pooling), a fully connected layer with 512 units and ReLu activation, and a final softmax output layer. This results in 1,663,370 parameters in total. We tune η from 0.01 to 0.5 with an increment value of 0.005. As in Kerkouche et al. [2020], we fix the momentum parameter ρ to 0.9 and the global learning rate η_G to 0.35. Same for the number of chunks used $P = 200$ (refers to Kerkouche et al. [2020] for more details). The hyperparameters used by each of the considered schemes are summarized in Table 5.

C COMPUTATIONAL ENVIRONMENT

Our experiments were performed on a server running Ubuntu 18.04 LTS equipped with a Intel(R) Xeon(R) Silver 4114 CPU @ 2.20GHz, 192GB RAM, and two NVIDIA Quadro P5000 GPU card of 16 Go each. We use Keras 2.2.0 Chollet et al. [2015a] with a TensorFlow backend 1.12.0 Abadi et al. [2015] and Numpy 1.14.3 Oliphant [2006] to implement our models and experiments. We use Python 3.6.5 and our code runs on a Docker container to simplify reproducibility.

D FURTHER EXPERIMENTS

The goal of this section is to compare the performance of our proposed schemes FL-TOP and FL-TOP-DP with several baselines according to different compression ratios. More specifically, we consider the following additional baselines:

- FL-BAS-2: As in FL-BASIC, only a randomly selected set of parameters are selected and sent to the server at each round. Importantly, none of the parameters are reinitialized during training.
- FL-BAS-3: This baseline is the same as FL-BASIC, except that the set of random parameters is fixed over all the rounds.
- FL-BAS-4: Same as FL-BAS-2, except that the set of random parameters is the same over all the rounds.
- FL-TOP-BIS: Similarly to FL-TOP, it uses the same Top-K parameters over the whole training. The only difference is that the $n-K$ non-Top-K parameters are not re-initialized after each SGD iteration. As in FL-TOP, after T_{gd} SGD iterations, clients send the update of the Top-K parameters to the server.

Note that all compression operators in the new baselines are still linear (just like FL-TOP-DP), and hence they can also be used with secure aggregation. Their private extensions (i.e., FL-BAS-2-DP, FL-BAS-3-DP, FL-BAS-4-DP and FL-TOP-BIS-DP) also clip and then noise the compressed updates as in FL-TOP-DP. The selection of sensitivity S happens similarly to FL-TOP-DP and FL-BASIC-DP using the public data as described in Section 4.

D.1 RESULTS

Table 8 shows the best accuracy over 200 rounds for each scheme on the Fashion-MNIST dataset. *Round* corresponds to the round when the best accuracy is achieved and *Cost* is the average bandwidth consumption calculated as: $r \times n \times 32 \times Round \times C$, where 32 is the number of bits necessary to represent a float value, n is the uncompressed model size, $r = \frac{|\mathbb{T}|}{n}$, $|\mathbb{T}|$ is the compressed model size, C is the sampling probability of a client, and *Round* is the round when we get the the best accuracy.

Table 9 and Table 10 display the best balanced accuracy over 100 rounds for each scheme on the Medical dataset. AUROC corresponds to the AUROC value when the best balanced accuracy is reached, *Round* is the round when we get the best balanced accuracy, and finally, Cost is the average bandwidth consumption calculated as for the Fashion-MNIST dataset described above.

On the medical data (see Table 9 and 10), our schemes FL-TOP and FL-TOP-DP reach 0.64 of balanced accuracy and 0.70 of AUROC for $r = 0.01\%$, while FL-TOP-Bis and

FL-TOP-Bis-DP, which are the best baselines, have 8% less of balanced accuracy and 10% less of AUROC for identical compression ratios. Furthermore, for larger compression ratios, FL-TOP and FL-TOP-DP have similar results to that of FL-TOP-Bis and FL-TOP-Bis-DP. However, above $r = 1\%$, FL-TOP outperforms FL-TOP-BIS. The same holds for FL-TOP-DP, which outperforms FL-TOP-Bis-DP when r is more than 0.05%.

On Fashion-MNIST, FL-TOP performs better than other schemes below $r = 10\%$. For $r = 10\%$, FL-CS and FL-TOP have the same accuracy of 0.85. FL-TOP-DP is the best DP scheme independently of the compression ratio r.

Notice the the larger the compression ratio r is the smaller the performance gap between our schemes and the baselines FL-BAS-1, FL-BAS-3. The same holds for their DP counterparts. This is mainly due to the fact that the larger r is the more likely that all schemes update the same Top-K parameters.

FL-CS and FL-CS-DP fail to improve their model accuracy when $r = 0.01\%$ on the medical dataset. The same holds for FL-BAS-3-DP when $r = 0.1\%$ on the Fashion-MNIST dataset.

On Fashion-MNIST, there is a decrease of accuracy for each of FL-TOP-DP, FL-TOP-BIS-DP and FL-CS-DP from $r = 5\%$ to $r = 10\%$. Indeed, as suggested in Kerkouche et al. [2020], it may be due to the increase of sensitivity S which will also increase the noise and therefore its negative impact on convergence.

Algorithm 4: FL-STD: Federated Learning

Server:
  Initialize common model w_0
  for $t = 1$ **to** T_{cl} **do**
    Select $\mathbb{K}$ clients uniformly at random
    for *each client k in* $\mathbb{K}$ **do**
      $\Delta \mathbf{w}_t^k = \mathbf{Client}_k(\mathbf{w}_{t-1})$
    end
    $\mathbf{w}_t = \mathbf{w}_{t-1} + \sum_k \frac{|D_k|}{\sum_j |D_j|} \Delta \mathbf{w}_t^k$
  end
  Output: Global model $\mathbf{w}_t$

Client$_k(\mathbf{w}_{t-1}^k)$**:**
  $\mathbf{w}_t^k = \mathbf{SGD}(D_k, \mathbf{w}_{t-1}^k, T_{\mathsf{gd}})$
  Output: Model update $(\mathbf{w}_t^k - \mathbf{w}_{t-1}^k)$

Algorithm 5: Stochastic Gradient Descent

Input: D : training data, T_{gd} : local epochs, $\mathbf{w}$: weights
for $t = 1$ **to** T_{gd} **do**
  Select batch $\mathbb{B}$ from D randomly
  $\mathbf{w} = \mathbf{w} - \eta \nabla f(\mathbb{B}; \mathbf{w})$
end
Output: Model $\mathbf{w}$

Algorithm 6: FL-STD-DP: Federated Learning with Client Privacy

Server:
  Initialize common model w_0
  for $t = 1$ **to** T_{cl} **do**
    Select $\mathbb{K}$ clients randomly
    for *each client k in* $\mathbb{K}$ **do**
      $\Delta \tilde{\mathbf{w}}_t^k = \mathbf{Client}_k(\mathbf{w}_{t-1})$
    end
    $\mathbf{w}_t = \mathbf{w}_{t-1} + \frac{1}{|\mathbb{K}|} \sum_k \Delta \tilde{\mathbf{w}}_t^k$
  end
Client$_k(\mathbf{w}_{t-1}^k)$**:**
  $\Delta \mathbf{w}_t^k = \mathbf{SGD}(D_k, \mathbf{w}_{t-1}^k, T_{\mathsf{gd}}) - \mathbf{w}_{t-1}^k$
  $\Delta \hat{\mathbf{w}}_t^k = \Delta \mathbf{w}_t^k / \max\left(1, \frac{||\Delta \mathbf{w}_t^k||_2}{S}\right)$
  Output: $\mathsf{Enc}_{K_k}(\mathscr{G}(\Delta \hat{\mathbf{w}}_t^k, S\mathbf{I}\sigma/\sqrt{|K|}))$

Table 3: Descriptions of features

Features	*Descriptions*
Age	Value in the range of 15 and 89
Gender	Male, Female or Unknown
Admission type	Emergency, Urgent, Trauma Center: visits to a trauma center/hospital or Unknown
MRCI	Medication regimen complexity index score (ranging from 2 to 60)
Drugs and ICD9 codes	Drugs given to the patient on the 1^{st} day of hospitalization. The ICD9 codes are composed of procedures and diagnosis codes, the first gives details about the medical procedures performed on the patient and the second about the doctor's diagnosis of the patient. There is a total of 24,419 possible drugs and ICD9 codes [CUADRADO, 2019].

Table 4: Number of instances for our case study. The Medical dataset contains in total 1,271,733 records.

Data	Positive cases	Negative cases	Ratio	Total
Train	32,106	985,280	3.16%	1,017,386
Test	7,882	246,465	3.10%	254,347

Datasets	*Common Parameters*
Fashion-MNIST dataset	$C = 1/60$; $N = 6000$; $T_{\text{cl}} = 200$; $T_{\text{gd}} = 5$; $\lvert\mathbb{B}\rvert = 10$; $\lvert D_k\rvert = 10$; $n = 1,663,370$; $\delta = 10^{-5}$; $SGD(\eta = 0.215)$; $\eta_G = 0.35$; $\rho = 0.9$; $P = 200$; $\sigma = 1.54$; $T_{\text{init}} = 5$
Medical dataset	$C = 100/5010$; $N = 5010$; $T_{\text{cl}} = 100$; $T_{\text{gd}} = 40$; $n = 1,496,601$; $\delta = 10^{-5}$; $SGD(\eta = 0.1)$; $\eta_G = 1.0$; $\rho = 0.9$; $P = 100$; $\sigma = 1.49$; $T_{\text{init}} = 40$

Table 5: Common environment between the schemes. ρ, η_G and P are only used with FL-CS and FL-CS-DP.

Algorithms	*Compression ratio (r)*				
	0.1%	*0.5%*	*1%*	*5%*	*10%*
FL-BASIC-DP	0.05	0.12	0.16	0.34	0.45
FL-BAS-2-DP	0.07	0.16	0.23	0.52	0.75
FL-BAS-3-DP	0.05	0.11	0.16	0.33	0.44
FL-BAS-4-DP	0.06	0.15	0.21	0.51	0.74
FL-CS-DP	0.21	0.26	0.32	0.57	0.79
FL-TOP-BIS-DP	1.25	1.59	1.79	2.18	2.34
FL-TOP-DP	0.50	0.61	0.64	0.87	1.0

Table 6: Sensitivity S used for each scheme and for different compression ratio r on Fashion-MNIST. For FL-STD-DP, S is set to 2.40.

Algorithms	*Compression ratio (r)*						
	0.01%	*0.05%*	*0.1%*	*0.5%*	*1%*	*5%*	*10%*
FL-BASIC-DP	0.01	0.03	0.05	0.11	0.16	0.34	0.46
FL-BAS-2-DP	0.01	0.03	0.04	0.09	0.14	0.31	0.44
FL-BAS-3-DP	0.01	0.04	0.06	0.12	0.18	0.35	0.49
FL-BAS-4-DP	0.02	0.03	0.05	0.12	0.15	0.31	0.44
FL-CS-DP	0.002	0.005	0.006	0.01	0.02	0.04	0.06
FL-TOP-BIS-DP	0.60	0.73	0.81	1.03	1.13	1.31	1.32
FL-TOP-DP	0.23	0.46	0.59	1.03	1.18	1.31	1.32

Table 7: Sensitivity S used for each scheme and for different compression ratio r on the medical dataset. For FL-STD-DP, S is set to 1.40.

Compression ratio (r)	Algorithms	*Performance*				
		Accuracy	*Round*	*Downstream Cost* (Kilobyte)	*Upstream Cost* (Kilobyte)	ε
0.1%	FL-BASIC	0.14	111	12308.94	12.31	N/A
	FL-BAS-2	0.16	185	20514.9	20.51	N/A
	FL-BAS-3	0.27	200	22.17	22.17	N/A
	FL-BAS-4	0.17	200	22.17	22.17	N/A
	FL-CS	0.37	200	22178.27	22.17	N/A
	FL-TOPK-BIS	0.59	198	21.95	21.95	N/A
	FL-TOP	**0.78**	199	**22.06**	**22.06**	N/A
	FL-BASIC-DP	0.14	167	18518.85	18.51	0.95
	FL-BAS-2-DP	0.14	124	13750.53	13.75	0.88
	FL-BAS-3-DP	-	-	-	-	-
	FL-BAS-4-DP	0.15	137	15.19	15.19	0.90
	FL-CS-DP	0.36	197	21845.59	21.84	1
	FL-TOPK-BIS-DP	0.59	196	21.73	21.73	0.99
	FL-TOP-DP	**0.76**	199	**22.06**	**22.06**	**1**
0.5%	FL-BASIC	0.65	193	21402.03	107	N/A
	FL-BAS-2	0.46	196	21734.70	108.66	N/A
	FL-BAS-3	0.73	200	110.88	110.88	N/A
	FL-BAS-4	0.41	197	109.22	109.22	N/A
	FL-CS	0.57	185	20514.9	102.56	N/A
	FL-TOPK-BIS	0.76	200	110.88	110.88	N/A
	FL-TOP	**0.82**	200	**110.88**	**110.88**	N/A
	FL-BASIC-DP	0.59	200	22178.27	110.88	1
	FL-BAS-2-DP	0.38	200	22178.27	110.88	1
	FL-BAS-3-DP	0.56	200	110.88	110.88	1
	FL-BAS-4-DP	0.33	200	110.88	110.88	1
	FL-CS-DP	0.53	200	22178.27	110.88	1
	FL-TOPK-BIS-DP	0.68	184	102.01	102.01	0.97
	FL-TOP-DP	**0.81**	200	**110.88**	**110.88**	**1**
1%	FL-BASIC	0.71	194	21512.92	215.12	N/A
	FL-BAS-2	0.59	200	22178.27	221.77	N/A
	FL-BAS-3	0.76	200	221.77	221.77	N/A
	FL-BAS-4	0.56	195	216.23	216.23	N/A
	FL-CS	0.69	200	22178.27	221.77	N/A
	FL-TOPK-BIS	0.79	197	218.45	218.45	N/A
	FL-TOP	**0.83**	200	**221.77**	**221.77**	N/A
	FL-BASIC-DP	0.65	197	21845.59	218.45	1
	FL-BAS-2-DP	0.62	198	21956.48	219.56	1
	FL-BAS-3-DP	0.66	198	219.56	219.56	1
	FL-BAS-4-DP	0.52	198	219.56	219.56	1
	FL-CS-DP	0.66	189	20958.46	209.58	0.98
	FL-TOPK-BIS-DP	0.70	174	192.94	192.94	0.96
	FL-TOP-DP	**0.81**	183	**202.92**	**202.92**	**0.97**
5%	FL-BASIC	0.78	196	21734.70	1086.73	N/A
	FL-BAS-2	0.72	199	22067.38	1103.36	N/A
	FL-BAS-3	0.81	199	1103.36	1103.36	N/A
	FL-BAS-4	0.76	196	1086.73	1086.73	N/A
	FL-CS	0.82	200	22178.27	1108.91	N/A
	FL-TOPK-BIS	0.83	196	1086.73	1086.73	N/A
	FL-TOP	**0.84**	200	**1108.91**	**1108.91**	N/A
	FL-BASIC-DP	0.76	195	21623.81	1081.18	0.99
	FL-BAS-2-DP	0.72	195	21623.81	1081.18	0.99
	FL-BAS-3-DP	0.76	199	1103.36	1103.36	1
	FL-BAS-4-DP	0.75	191	1059.01	1059.01	0.99
	FL-CS-DP	0.78	160	17742.61	887.13	0.94
	FL-TOPK-BIS-DP	0.71	152	842.77	842.77	0.92
	FL-TOP-DP	**0.81**	152	**842.77**	**842.77**	**0.92**
10%	FL-BASIC	0.81	196	21734.70	2173.47	N/A
	FL-BAS-2	0.78	199	22067.38	2206.74	N/A
	FL-BAS-3	0.82	195	2162.38	2162.38	N/A
	FL-BAS-4	0.79	200	2217.83	2217.83	N/A
	FL-CS	0.85	182	20182.22	2018.22	N/A
	FL-TOPK-BIS	0.84	196	2173.47	2173.47	N/A
	FL-TOP	**0.85**	199	**2206.74**	**2206.74**	N/A
	FL-BASIC-DP	0.79	189	20958.46	2095.85	0.98
	FL-BAS-2-DP	0.77	189	20958.46	2095.85	0.98
	FL-BAS-3-DP	0.79	183	2029.31	2029.31	0.97
	FL-BAS-4-DP	0.78	195	2162.38	2162.38	0.99
	FL-CS-DP	0.72	167	18518.85	1851.89	0.95
	FL-TOPK-BIS-DP	0.69	138	1530.30	1530.30	0.90
	FL-TOP-DP	**0.80**	157	**1740.99**	**1740.99**	**0.93**
100%	FL-STD	0.86	200	22178.27	22178.27	N/A
	FL-STD-DP	0.56	60	6653.48	6653.48	0.76

Table 8: Summary of results on Fashion-MNIST dataset.

Compression ratio (r)	Algorithms	*Performance*					
		Bal_Acc	*AUROC*	*Round*	*Downstream Cost* (Kilobyte)	*Upstream Cost* (Kilobyte)	ε
0.01%	FL-BASIC	0.49	0.45	100	11948.91	1.19	N/A
	FL-BAS-2	0.49	0.45	94	11231.98	1.12	N/A
	FL-BAS-3	0.49	0.45	81	0.96	0.96	N/A
	FL-BAS-4	0.49	0.49	100	1.19	1.19	N/A
	FL-CS	-	-	-	-	-	N/A
	FL-TOP-Bis	0.59	0.63	100	1.19	1.19	N/A
	FL-TOP	**0.64**	**0.70**	60	**0.71**	**0.71**	N/A
	FL-BASIC-DP	0.49	0.45	6	716.93	0.07	0.74
	FL-BAS-2-DP	0.49	0.45	100	11948.91	1.19	1
	FL-BAS-3-DP	0.49	0.45	95	1.13	1.13	0.99
	FL-BAS-4-DP	0.49	0.47	96	1.14	1.14	0.99
	FL-CS-DP	-	-	-	-	-	-
	FL-TOP-Bis-DP	0.59	0.63	94	1.12	1.12	0.99
	FL-TOP-DP	**0.64**	**0.70**	100	**1.19**	**1.19**	**1**
0.05%	FL-BASIC	0.50	0.48	100	11948.91	5.97	N/A
	FL-BAS-2	0.49	0.46	100	11948.91	5.97	N/A
	FL-BAS-3	0.51	0.49	100	5.97	5.97	N/A
	FL-BAS-4	0.51	0.52	57	3.40	3.40	N/A
	FL-CS	0.51	0.50	100	11948.91	5.97	N/A
	FL-TOP-Bis	0.68	0.75	92	5.49	5.49	N/A
	FL-TOP	**0.68**	**0.75**	54	**3.22**	**3.22**	N/A
	FL-BASIC-DP	0.49	0.46	84	10037.08	5.02	0.96
	FL-BAS-2-DP	0.49	0.46	100	11948.91	5.97	1
	FL-BAS-3-DP	0.50	0.48	99	5.91	5.91	1
	FL-BAS-4-DP	0.52	0.51	100	5.97	5.97	1
	FL-CS-DP	0.49	0.48	100	11948.91	5.97	1
	FL-TOP-Bis-DP	0.68	0.75	92	5.49	5.49	0.98
	FL-TOP-DP	**0.68**	**0.75**	99	**5.91**	**5.91**	**1**
0.1%	FL-BASIC	0.51	0.51	99	11829.42	11.82	N/A
	FL-BAS-2	0.50	0.47	100	11948.91	11.94	N/A
	FL-BAS-3	0.53	0.53	100	11.94	11.94	N/A
	FL-BAS-4	0.50	0.53	94	11.23	11.23	N/A
	FL-CS	0.53	0.55	100	11948.91	11.94	N/A
	FL-TOP-Bis	0.69	0.76	100	11.94	11.94	N/A
	FL-TOP	**0.69**	**0.76**	68	**8.12**	**8.12**	N/A
	FL-BASIC-DP	0.50	0.49	100	11948.91	11.94	1
	FL-BAS-2-DP	0.50	0.47	100	11948.91	11.94	1
	FL-BAS-3-DP	0.55	0.56	100	11.94	11.94	1
	FL-BAS-4-DP	0.51	0.52	100	11.94	11.94	1
	FL-CS-DP	0.51	0.51	99	11829.42	11.82	1
	FL-TOP-Bis-DP	0.68	0.75	89	10.63	10.63	0.98
	FL-TOP-DP	**0.69**	**0.76**	85	**10.15**	**10.15**	**0.97**
0.5%	FL-BASIC	0.58	0.68	100	11948.91	59.74	N/A
	FL-BAS-2	0.56	0.58	99	11829.42	59.15	N/A
	FL-BAS-3	0.61	0.68	100	59.74	59.74	N/A
	FL-BAS-4	0.56	0.59	100	59.74	59.74	N/A
	FL-CS	0.66	0.71	100	11948.91	59.74	N/A
	FL-TOP-Bis	0.71	0.78	100	59.74	59.74	N/A
	FL-TOP	**0.71**	**0.79**	95	**56.76**	**56.76**	N/A
	FL-BASIC-DP	0.57	0.64	100	11948.91	59.74	1
	FL-BAS-2-DP	0.57	0.59	100	11948.91	59.74	1
	FL-BAS-3-DP	0.58	0.67	100	59.74	59.74	1
	FL-BAS-4-DP	0.54	0.57	34	20.31	20.31	0.83
	FL-CS-DP	0.61	0.68	100	11948.91	59.74	1
	FL-TOP-Bis-DP	0.68	0.75	55	32.86	32.86	0.89
	FL-TOP-DP	**0.69**	**0.76**	24	**14.34**	**14.34**	**0.80**

Table 9: Summary of results on Medical dataset (Part 1).

Compression ratio (r)	Algorithms	*Performance*					
		Bal_Acc	*AUROC*	*Round*	*Downstream Cost* (Kilobyte)	*Upstream Cost* (Kilobyte)	ε
1%	FL-BASIC	0.64	0.72	100	11948.91	119.49	N/A
	FL-BAS-2	0.62	0.66	100	11948.91	119.49	N/A
	FL-BAS-3	0.62	0.66	85	101.57	101.57	N/A
	FL-BAS-4	0.56	0.59	100	119.49	119.49	N/A
	FL-CS	0.68	0.75	100	11948.91	119.49	N/A
	FL-TOP-Bis	0.72	0.79	100	119.49	119.49	N/A
	FL-TOP	**0.72**	**0.79**	58	**69.30**	**69.30**	N/A
	FL-BASIC-DP	0.64	0.70	100	11948.91	119.49	1
	FL-BAS-2-DP	0.62	0.67	100	11948.91	119.49	1
	FL-BAS-3-DP	0.61	0.71	100	119.49	119.49	1
	FL-BAS-4-DP	0.57	0.66	100	119.49	119.49	1
	FL-CS-DP	0.66	0.72	100	11948.91	119.49	1
	FL-TOP-Bis-DP	0.68	0.74	53	63.33	63.33	0.89
	FL-TOP-DP	**0.69**	**0.76**	22	**26.29**	**26.29**	**0.79**
5%	FL-BASIC	0.72	0.80	100	11948.91	597.45	N/A
	FL-BAS-2	0.68	0.75	100	11948.91	597.45	N/A
	FL-BAS-3	0.69	0.76	98	585.5	585.5	N/A
	FL-BAS-4	0.66	0.72	100	597.45	597.45	N/A
	FL-CS	0.73	0.81	98	11709.93	585.5	N/A
	FL-TOP-Bis	0.72	0.79	100	597.45	597.45	N/A
	FL-TOP	**0.72**	**0.80**	95	**567.57**	**567.57**	N/A
	FL-BASIC-DP	0.69	0.76	100	11948.91	597.45	1
	FL-BAS-2-DP	0.68	0.75	98	11709.93	585.5	1
	FL-BAS-3-DP	0.65	0.71	90	537.70	537.70	0.98
	FL-BAS-4-DP	0.67	0.74	98	585.5	585.5	1
	FL-CS-DP	0.69	0.76	100	11948.91	597.45	1
	FL-TOP-Bis-DP	0.67	0.74	38	227.03	227.03	0.84
	FL-TOP-DP	**0.68**	**0.75**	23	**137.41**	**137.41**	**0.79**
10%	FL-BASIC	0.74	0.81	100	11948.91	1194.89	N/A
	FL-BAS-2	0.70	0.77	100	11948.91	1194.89	N/A
	FL-BAS-3	0.72	0.80	98	1170.99	1170.99	N/A
	FL-BAS-4	0.70	0.77	99	1182.94	1182.94	N/A
	FL-CS	0.74	0.82	100	11948.91	1194.89	N/A
	FL-TOP-Bis	0.72	0.80	100	1194.89	1194.89	N/A
	FL-TOP	**0.74**	**0.82**	90	**1075.40**	**1075.40**	N/A
	FL-BASIC-DP	0.69	0.76	99	11829.42	1182.94	1
	FL-BAS-2-DP	0.69	0.76	95	11351.46	1135.15	0.99
	FL-BAS-3-DP	0.69	0.76	95	1135.15	1135.15	0.99
	FL-BAS-4-DP	0.69	0.76	100	1194.89	1194.89	1
	FL-CS-DP	0.69	0.76	96	11470.95	1147.09	0.99
	FL-TOP-Bis-DP	0.67	0.73	37	442.11	442.11	0.84
	FL-TOP-DP	**0.68**	**0.74**	23	**274.82**	**274.82**	**0.79**
100%	FL-STD	0.74	0.82	99	11829.42	11829.42	N/A
	FL-STD-DP	0.66	0.72	62	7408.32	7408.32	0.91

Table 10: Summary of results on Medical dataset (Part 2).