Constrained Differentially Private Federated Learning for Low-bandwidth Devices (Supplementary material)

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

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[1] 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 1. 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

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

Supplement for the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2021).

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 2 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 1. 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 [2]. 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[3]). 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 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 2).

   Using Client-Level differential privacy requires adding more noise than Record-Level differential privacy, because the privacy purposes are not the same as detailed 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. [4]

A.4 PERFORMANCE METRICS

We use the following metrics:

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

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

[4] 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.

- *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 2), 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 3.

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 3.

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 6 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 7 and Table 8 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 7 and 8), 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.

Table 1: 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 2: 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

Algorithm 1: 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 $\mathbf{w}_t$

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

References

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

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

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.

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.

Algorithm 2: 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 3: 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
$\mathbf{Client}_k(\mathbf{w}_{t-1}^k)$**:**
  $\Delta\mathbf{w}_t^k = \mathbf{SGD}(D_k, \mathbf{w}_{t-1}^k, T_{\mathsf{gd}}) - \mathbf{w}_{t-1}^k$
  $\Delta\hat{\mathbf{w}}_t^k = \Delta\mathbf{w}_t^k / \max\left(1, \frac{||\Delta\mathbf{w}_t^k||_2}{S}\right)$
  Output: $\mathsf{Enc}_{K_k}(\mathcal{G}(\Delta\hat{\mathbf{w}}_t^k, S\mathbf{I}\sigma/\sqrt{|K|}))$

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

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

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.

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

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

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

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

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

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.

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

Table 3: 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 4: 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 5: 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.

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.

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

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.

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

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 6: 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 7: 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 8: Summary of results on Medical dataset (Part 2).