Heliyon 9 (2023) e16925

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Benchmarking federated strategies in Peer-to-Peer Federated learning for biomedical data

Jose L. Salmeron [a,b], Irina Arévalo [c,*], Antonio Ruiz-Celma [d]

[a] *CUNEF Universidad, Madrid, Spain*
[b] *Universidad Autónoma de Chile, Chile*
[c] *Universidad Pablo de Olavide, Seville, Spain*
[d] *Universidad de Extremadura, Badajoz, Spain*

ARTICLE INFO

Keywords:
Federated learning
Privacy-preserving machine learning

ABSTRACT

The increasing requirements for data protection and privacy have attracted a huge research interest on distributed artificial intelligence and specifically on federated learning, an emerging machine learning approach that allows the construction of a model between several participants who hold their own private data. In the initial proposal of federated learning the architecture was centralised and the aggregation was done with federated averaging, meaning that a central server will orchestrate the federation using the most straightforward averaging strategy. This research is focused on testing different federated strategies in a peer-to-peer environment. The authors propose various aggregation strategies for federated learning, including weighted averaging aggregation, using different factors and strategies based on participant contribution. The strategies are tested with varying data sizes to identify the most robust ones. This research tests the strategies with several biomedical datasets and the results of the experiments show that the accuracy-based weighted average outperforms the classical federated averaging method.

1. Introduction

Artificial intelligence applications in healthcare are increasing every day. These applications have the ability to advance the healthcare industry by, for instance, supporting clinical decision making, risk prediction, developing early warning systems for patients, increasing the accuracy and timeliness of diagnosis, improving patient–physician interaction, and optimizing operations and resource allocation [21].

Federated learning is a new approach for distributed artificial intelligence that aims to have several agents train a deep learning model in a collaborative and secure way, without sharing any private data. This training is done the following way: a central server defines a deep learning model and sends it to the agents, who train the model in their private data. Then, they send the parameters of the model (weights or gradients) back to the server, who aggregates these data in order to find a global federated model, which in turn is delivered back to the agents to be retrained in their data. This process is iterated until convergence.

In the initial definition of the federated learning approach, the aggregation step is done by averaging the model parameters. Nevertheless, other aggregation methods may be of more interest since they can improve the performance of the model by giving more weight to different agents depending on their size or the performance of the local models in their data.

The main contributions of this research are two-fold:

[*] Corresponding author.
E-mail addresses: joseluis.salmeron@cunef.edu (J.L. Salmeron), iarebar@alu.upo.es (I. Arévalo), aruiz@unex.es (A. Ruiz-Celma).

https://doi.org/10.1016/j.heliyon.2023.e16925
Received 5 December 2022; Received in revised form 31 May 2023; Accepted 1 June 2023

Available online 2 June 2023

1. Several aggregation strategies are proposed, such as weighted averaging aggregation using the dataset size, weighted average using the normalized inverse of the local test accuracy, weighted averaging aggregation using the dataset size and accuracy, weighted average using the contribution of the participant, and weighted sum using the inverse contribution of the participant. Federated averaging is included for comparison.
2. The strategies are tested with different data sizes on each participant. This allows analyzing the strategies under different circumstances and identifying those that are more robust.

The rest of this paper is organized as follows. We discuss the theoretical background of federated learning in section 2. The different federated strategies are described in section 2.3. The methodological proposal can be found in section 3, while section 4 shows the results of the experimental approach, serving as a benchmark. Finally, the authors draw a conclusion in section 5.

2. Fundamentals

2.1. Related work

Federated learning is an emerging approach for distributed artificial intelligence in which the different data owners (or participants) train collaboratively a machine learning model [17,22]. The model is updated (trained) in the own private data of each participant and then the trained model is sent for aggregation to a central server or one of the participants. It was first proposed by McMahan et al. [18] and further developed in Konecny et al. [14] and McMahan and Ramage [19]. The main advantage of federated learning is the training of a model with the private data of each participant keeping the security and compliance requirements while improving their models [3]. It also allows the use of more accurate models with low latency, ensuring privacy and less power consumption [31].

Numerous surveys and literature reviews have extensively examined the body of work documented in academic literature pertaining to architectures, approaches, utilization, and applications of federated learning. In the healthcare domain, Hoyos et al. [10] present federated learning approaches (horizontal, vertical and transfer learning) for FCMs for the prediction of mortality and the prescription of treatment of severe dengue. Antunes et al. [4] outline a broad architecture for federated learning applied to healthcare data, drawing upon key insights derived from the literature review. Li et al. [15] analyse recent literature on the utilization of federated learning, outlining various federated learning architectures and classification models. Nguyen et al. [20] provide a comprehensive and up-to-date review of the latest advancements in federated learning within crucial healthcare domains, encompassing health data management, remote health monitoring, medical imaging, and COVID-19 detection. Xu et al. [30] provide an overview of the common solutions to address statistical challenges, system challenges, and privacy issues in federated learning. They also highlight the potential implications and opportunities that federated learning holds for the healthcare sector. Furthermore, numerous studies have sought to optimize federated learning and broaden its practical applications, with research efforts spanning areas such as computation fusion [32], data transmission [23,26], as well as privacy and security-related concerns [9].

2.2. Architecture

Fully decentralised learning aims to replace server-based communication with peer-to-peer communication among individual clients as its core concept. Each round in fully decentralized algorithms involves a client performing a local update and sharing information with their neighbours in the graph. In this research, clients share the model with all participants, making it a fully connected peer-to-peer architecture (see Fig. 1).

Peer-to-peer architectures can have a significant impact on federated learning. With peer-to-peer communication, clients can collaborate and share locally trained models directly with each other, which can result in faster and more efficient learning compared to a centralized architecture. Peer-to-peer architectures can also provide better privacy and security as the clients can keep their data locally and only share the necessary information with their peers.

Recently, a fully decentralised solution where participants collaborate asynchronously and communicate in a peer-to-peer fashion, without any central server to orchestrate a global state of the system or even to coordinate the protocol, is proposed in [28]. In this scenario, peer-to-peer architectures can also have some challenges. For example, it can be more difficult to coordinate and manage the communication between clients, and it may require additional mechanisms to ensure the consistency of the model across all clients. To overcome this challenge, this paper focuses on a fully connected peer-to-peer architecture. Moreover, this kind of systems may not be suitable for large-scale federated learning scenarios due to the high communication overhead between clients. However, the goal of this research is to present a fault-tolerant architecture in case of failure of a central server and/or in multiple participants.

In this case, the group of autonomous peers run iteratively multiple training rounds to update the federated model [28,29]. Indeed, peer-to-peer algorithms include scalability-by-design to large sets of devices thanks to the locality of their updates [13]. In addition, a decentralised peer-to-peer architecture intrinsically provides an additional some security guarantee as it becomes much more tough for any third party to get the full state information of the system [29].

A peer-to-peer federated learning process needs a minimum of two participants. In this case, there is not a central server managing the federated model and the communications with the participants. In a peer-to-peer architecture, each participant receives the trained models of the remainder participants and then averages the participants' models to obtain a federated model trained with the private data. The participants own the data and train the partial models. This process can be iterated as many times as needed. The process is shown in Fig. 1 and is as follows:

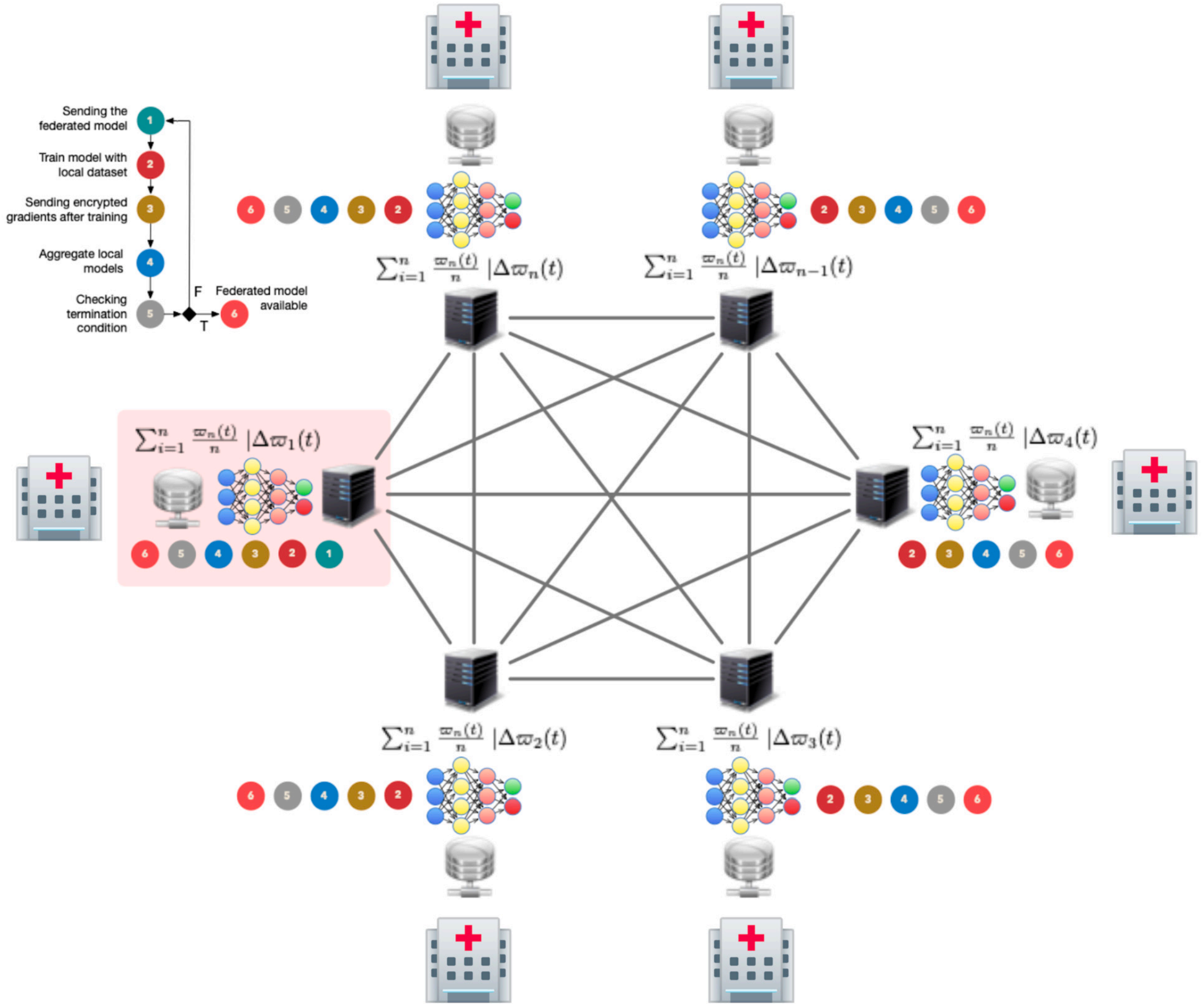
Sending the federated model
1
Train model with local dataset
2
Sending encrypted gradients after training
3
Aggregate local models
4
Checking termination condition
5
F
Federated model available
6
T
$\sum_{i=1}^{n} \frac{\varpi_n(t)}{n} |\Delta\varpi_n(t)$
$\sum_{i=1}^{n} \frac{\varpi_n(t)}{n} |\Delta\varpi_{n-1}(t)$
$\sum_{i=1}^{n} \frac{\varpi_n(t)}{n} |\Delta\varpi_1(t)$
$\sum_{i=1}^{n} \frac{\varpi_n(t)}{n} |\Delta\varpi_4(t)$
$\sum_{i=1}^{n} \frac{\varpi_n(t)}{n} |\Delta\varpi_2(t)$
$\sum_{i=1}^{n} \frac{\varpi_n(t)}{n} |\Delta\varpi_3(t)$

Fig. 1. Fully connected peer-to-peer federated learning architecture.

1. A participant initiates the federated learning process by sending an initial model to all the other participants. If this is the initial iteration, the federated model is dispatched by the participant triggering the process.
2. Each participant trains the received model using their own private data.
3. Each participant sends the parameters of the model in a private way (usually encrypting the data to be sent) to the remaining participants.
4. Every participant aggregates the partial models with the updated parameters and builds the federated model according to a federation strategy.
5. Every participant checks a termination condition in either accuracy of the model in a test dataset or number of iterations. If it is accomplished, the federated learning process ends, otherwise the process iterate again from step 1.

The target of the federated learning process is to minimize the total loss for all participants, computed as in Equation (1),

$$\mathcal{L}^* = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_i(\mathcal{D}_i, \Phi) \tag{1}$$

where $\mathcal{L}^*$ is the loss function for the federated model, $\mathcal{L}_i$ is the loss function for each participant in the federation, $\mathcal{D}_i$ is the dataset of the participant i and Φ is the federated model parameters, n is the number of participants.

The federated learning process trains a model between different participants without the sharing of private data. Nevertheless, there are other possible risks, like model poisoning, potential attacks to reconstruct the model or the training data from the parameters that the participants send to the central server, or the use of attack models [5,27]. As a consequence, there have been several advances in the use of privacy-preserving methods such as Differential Privacy or Homomorphic Encryption in federated learning, see [1,2,12,11,6].

Algorithm 1: Federated learning.

```
Data: The $K$ clients are indexed by $k$; $B$ is the local mini-batch size, $E$ is the number of local epochs, $T$ is the maximum iteration number, and $\eta$ is the
      learning rate.
Federation process
initialize $\Phi_0$
for each round $t = 1, 2, \dots, T$ do
    $m \leftarrow \max(C \cdot K, 1)$
    for each client $k \in S_t \mid S_t \sim U(m)$ in parallel do
        $\Phi^k_{t+1} \leftarrow client\_update(k, \Phi_t)$
    end
    $\Phi_{t+1} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} \cdot \Phi^k_{t+1}$
end
Training process
client_update($k, \Phi$): $B \leftarrow$(split $\mathcal{P}_k$ into batches of size $B$);
for each local epoch $i$ from 1 to $E$ do
    for batch $b \in B$ do
        $w \leftarrow \Phi - \eta \nabla l(\Phi; b)$
    end
end
return $\Phi$ to server
```

2.3. Federation strategies

As mentioned before, the aggregation method for federated learning is an important parameter of the process. The original definition contemplated the arithmetic mean of the parameters of the model to obtain the federated model.

In this research, the authors propose a series of different aggregations and compare them in our experimental section (see section 4). Assuming that the parameters of the model at iteration j are as shown in Equation (2),

$$\Phi_j = [\Phi_{j1}, \Phi_{j2}, \cdots, \Phi_{jn}] \tag{2}$$

where n is the number of participants, $\mathcal{D}_i$ is the dataset of the participant i, and Φ'_j is the parameters (weights or gradients) of the federated model, the functions of the parameters that we will discuss are the following:

- Average of the parameters (weights or gradients):

  $$\Phi'_j = \frac{1}{n} \sum_{i=1}^{n} \Phi_{ji} \tag{3}$$

  where every participant contributes the same to the global model.
- Weighted averaging aggregation using the normalized size of each participants' dataset:

  $$\Phi'_j = \sum_{i=1}^{n} \frac{|\mathcal{D}_i|}{\sum_{k=1}^{n} |\mathcal{D}_k|} \cdot \Phi_{ji} \tag{4}$$

  where every participant contributes to the global model proportionally to the size of their data, and agents with less information will affect less to the final model.
- Weighted average using the normalized inverse accuracy of the model in a test set of each participant:

  $$\Phi'_j = \sum_{i=1}^{n} \frac{1/\mathrm{acc}_{ji}}{\sum_{k=1}^{n} 1/\mathrm{acc}_{jk}} \cdot \Phi_{ji} \tag{5}$$

  where the individual models add to the global model inversely to their performance metric, trying to give more weight to the less accurate models in order to improve their metric in their datasets.
- Weighted average using the accuracy and the size of the dataset:

  $$\Phi'_j = \sum_{i=1}^{n} \frac{\mathrm{acc}_{ji} |\mathcal{D}_i|}{\sum_{k=1}^{n} |\mathcal{D}_k|} \cdot \Phi_{ji} \tag{6}$$

  where the contribution of each participant's model depends on both the accuracy of the model and the size of the dataset.
- Weighted average using the contribution ($\mathcal{C}$) of the participant, that is, the normalized inverse of the absolute difference between the loss of the participant's model and the loss of the global model when applied to the participants' data as shown in Equation (7)

$$C_{ji} = |\mathcal{L}^*_j(\mathcal{D}_i, \Phi) - \mathcal{L}_j(\mathcal{D}_i, \Phi)| \tag{7}$$

and

$$\Phi'_j = \sum_{i=1}^{n} \frac{C_{j,i}}{\sum_{k=1}^{n} C_{jk}} \cdot \Phi_{ji} \tag{8}$$

- Weighted sum using the inverse contribution of the participant:

$$\Phi'_j = \sum_{i=1}^{n} \frac{1/C_{ji}}{\sum_{k=1}^{n} 1/C_{jk}} \cdot \Phi_{ji} \tag{9}$$

3. Methodological proposal

The federated learning proposal starts with a central server sending an untrained model to the participants. As a first step, each hospital trains this model with their training data, evaluates it in their test data, and sends the parameters of the trained model back to the server. After receiving all the parameters from all the participants, in this proposal the server aggregates the parameters using one of the aggregation methods described in Section 2.3 to obtain the global method. This process is iterated until convergence. In this use case the authors have proposed a peer-to-peer architecture, and have not included an additional encryption layer, but it is possible and desirable to do so in real-life applications in healthcare.

In the following experiments, the initial model will be a dense neural network made of five layers followed by a non-linear ReLU function and a dropout layer for regularization (Fig. 2). For training, the loss function will be computed using the Binary Cross Entropy.

As a setup for the experiments, the authors assume that several hospitals with their own private data wish to train a deep learning model for diagnosis of a disease, but the size of their data is not large enough for training an accurate model. The data of all the hospitals should not be combined due to data regulations given their special sensitivity. Therefore, for each one of the experiments the corresponding dataset will be randomly split in five different subsets, to simulate five hospitals. In a first test all of them will have the same amount of data, obtained by splitting evenly the dataset. The other three tests will be random splits among the participants, where in the last two we have forced that there would be several participants with a very small number of samples (less than 10%). The distribution of the variables, including the percentage of positive cases for the target, will also vary from one hospital to another.

In every case, each participant's data will be split into a train and a test dataset for training and evaluation.

The metric we will use to compare the performance of the different models and methods is the accuracy on the test set. Given a classification model, its performance can be summarized with a confusion matrix, a table that shows the number of real positive and negative samples in our dataset versus the number of positive and negative values that our classifier predicts.

The accuracy of a model is the ratio of number of correct predictions to the total number of input samples, that is, the sum of the main diagonal of the confusion matrix divided by the sum of all values in the table, which is the amount of times the classifier got the prediction right. This metric varies between 0 and 1, where 0 is the worst case scenario, and 1 is associated to a perfect classifier. Moreover, a random binary classifier such as a coin toss has an accuracy of 0.5, and therefore this is a first baseline for any balanced binary classifier, understanding that an accuracy close to 0.5 is as bad as a random guess.

The results of the experiments will be shown as follows: for each experiment, each line will represent one of the five participant hospitals. The columns represent the number of the partition, its size, and the percentage of positives in the partition, and the accuracy of a model evaluated in each participant's test set. The first one is the local model in every participant without any federation, and the following are the results for each aggregation method described in Section 2.3: Federated Average (Equation (3)) and weighted sums with size (Eq. (4)), inverse accuracy (Eq. (5)), size and accuracy (Eq. (6)), contribution (Eq. (8)), and inverse contribution (Eq. (9)). The average accuracy for all participants in each experiment will also be included and used for comparison of the aggregation methods.

Table 1 summarizes how the results of the experiments will be shown, and what federation strategies are being evaluated.

These results will serve as a benchmark for which averaging strategy improves the most the performance of the federated learning process, by comparing the accuracy in different datasets of the competing approaches. The benchmarking methodology is as follows:

1. Define the problem: Firstly, define the problem that the AI model is intended to solve. In this research, we have selected four medical problems (breast cancer, chronic kidney disease, Parkinson's disease, and heart disease) to better validate our proposals.
2. Identify the relevant data: We have selected well-known datasets for each of these medical problems to ensure reproducibility.
3. Preprocess and split the data: The datasets have already been preprocessed and are ready for use. The next step is to split the data into training and testing sets to evaluate the performance of the model. The training set will be used to train the model, while the testing set will be used to measure its accuracy on unseen data.
4. Train the model: Train the Federated FCM model using different aggregation strategies.
5. Evaluate the model: Evaluate the performance of the model using the test set using the accuracy metric.
6. Compare the results: Conduct a comparison of the outcomes achieved by each aggregation strategy with the remaining ones to ascertain whether there are any significant performance differences among the strategies.

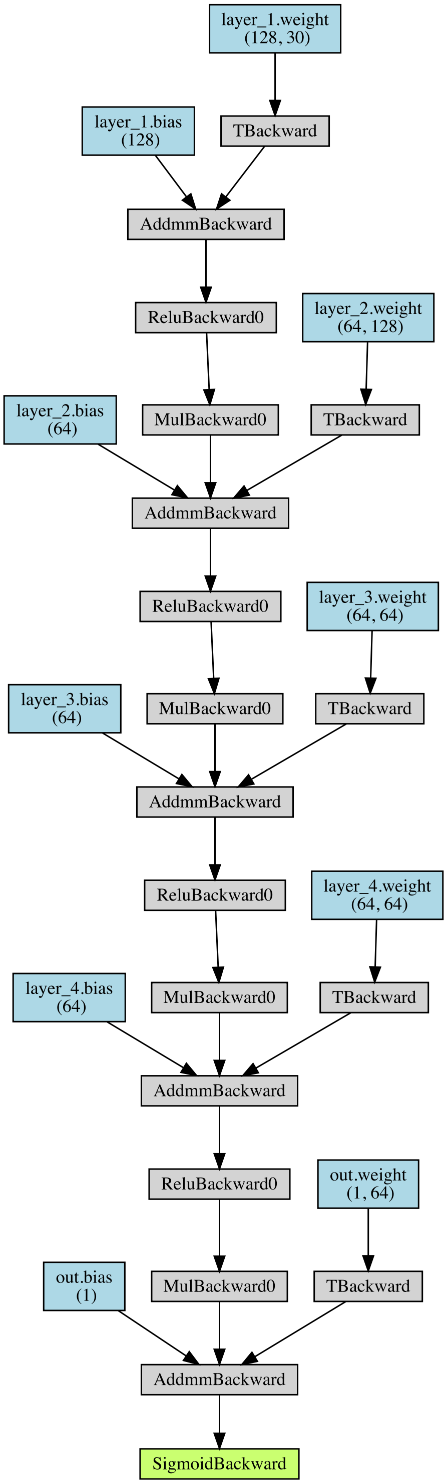

Fig. 2. Deep Neural Network topology for the experiments.

Table 1
Column names and federation strategies.

Column name	Definition	Federation strategy
Part	Partition number	
Size	Percentage of the total samples in this agent	
Pos (%)	Percentage of positives in the partition	
Acc (local)	Accuracy in the non-federated case	
Acc Federated Averaging (Fed)	Accuracy of the federated model using fed averaging	Equation (3)
Acc. Size (Fed)	Accuracy of the model using size-based averaging	Equation (4)
Acc. Accuracy (Fed)	Accuracy of the federated model using accuracy-based averaging	Equation (5)
Acc. Size & Acc. (Fed)	Accuracy of the federated model using both accuracy and size in the averaging	Equation (6)
Acc. Cont (Fed)	Accuracy of the federated model using the contribution of the participants	Equation (8)
Acc. Inv. Cont. (Fed)	Accuracy of the federated model using the inverse contribution of the participants	Equation (9)

4. Experiments

4.1. Experiment 1 - Breast Cancer

The Breast Cancer Wisconsin Dataset contains descriptions of features of the nucleus of a breast mass, obtained from digitized images of the fine-needle aspirate, for 569 patients, where 212 are malignant and 357 are benign tumors. This dataset is publicly available [8]. More details can be found in [24], [25].

The results of the federated learning process for this dataset are shown in Table 2. As we can see, the federation improves the accuracy metrics in general, but there are differences between the aggregation methods used. In the first case of the even dataset all accuracies are improved, but in the first uneven split, the contribution aggregation of the difference of the losses does not induce an increase in the performance. In the second uneven split, more extreme than the previous one, with one participant with only 3% of the data, the Federated Averaging does not improve the performance of the local models, and in the last one, with two participants with 6% and 7% of the data, the weighted aggregation using the size of each participant and the size and the accuracy show a lower accuracy than the first iteration of local models.

Summing up, for this experiment the accuracy-based aggregation and the inverse contribution are the only aggregation methods that improve the performance of the local models after the federation process.

4.2. Experiment 2 - chronic kidney disease

The term chronic kidney disease (CKD) describes all degrees of decreased renal function. It is more prevalent in the elderly population and it is estimated that affects 10–15% of the world population. CKD is not often identified in premature stages.

The Chronic Kidney Dataset contains 25 features and a target that represents whether the patient has the Chronic Kidney Disease, for 400 patients, one third of which did not have the disease and two thirds that did. It is publicly available at the UC Irvine Machine Learning Repository [8].

The results of the experiments for this dataset are shown in Table 3. In this case, the aggregation using the inverse contribution does not improve the accuracy for all participants, since this metric worsens for the first uneven split. The size-based and contribution-base aggregation do not increase the accuracy of the local models for this split as well, while the Federated Averaging fails for the last uneven split and the size and accuracy weighted average for the second one. As in the previous experiment, the only aggregation method that impoves the accuracy for all cases is the accuracy-based one.

4.3. Experiment 3 - Parkinson's

Parkinson's is a neurodegenerative disease that produces alterations in gait and posture that may increase the risk of falls and leads to mobility disabilities. Parkinson's affects about 1% of the world population over the age of 55.

The symptoms generally develop over years and their progressions are very diverse, making the diagnosis of the disease in the early stages extremely difficult.

Table 2
Experiment 1 - Breast Cancer.

Part	Size	Pos (%)	Acc (local)	Acc Federated Averaging (Fed)	Acc Size (Fed)	Acc Accuracy (Fed)	Acc Size & acc (Fed)	Acc Cont (Fed)	Acc Inv Cont (Fed)
1	20%	47%	0.5454	0.6363	0.6363	0.3636	0.7272	0.6363	0.7272
2	20%	41%	0.4166	0.5000	0.5833	0.7500	0.7500	0.6666	0.7500
3	20%	31%	0.5454	0.7272	0.8181	0.5454	0.8181	0.6363	0.8181
4	20%	35%	0.5833	0.5833	0.6666	0.5833	0.7500	0.4166	0.8333
5	20%	32%	0.5833	0.6666	0.6666	0.7500	0.4166	0.5833	0.5833
Avg	–	–	0.5348	**0.6227**	**0.6742**	**0.5984**	**0.6924**	**0.5879**	**0.7424**
1	20%	12%	0.9090	0.9090	0.8000	0.8181	0.8888	0.8181	0.8181
2	22%	30%	0.6153	0.5384	0.7692	0.5333	0.5714	0.7058	0.5000
3	18%	47%	0.5000	0.6000	0.3636	0.4444	0.5454	0.4285	0.4000
4	17%	17%	0.7000	0.7000	0.9000	0.7500	0.7500	0.6363	0.8000
5	24%	39%	0.4285	0.5714	0.6666	0.6666	0.4285	0.3571	0.7142
Avg	–	–	0.6306	**0.6637**	**0.6999**	**0.6425**	**0.6368**	0.5892	**0.6465**
1	49%	28%	0.5714	0.6428	0.6333	0.6896	0.7142	0.7096	0.6896
2	3%	40%	0.5000	0.0000	1.0000	0.6666	1.0000	0.5000	0.6666
3	15%	40%	0.8888	1.0000	0.5000	0.8333	0.6250	0.5000	0.8333
4	5%	63%	0.3333	0.3333	1.0000	0.6666	0.0000	0.6666	0.6666
5	29%	20%	0.4117	0.5882	0.7500	0.5789	0.7368	0.8666	0.5789
Avg	–	–	0.5410	0.5128	**0.7766**	**0.6870**	**0.6152**	**0.6486**	**0.6870**
1	48%	30%	0.6666	0.7333	0.7419	0.6969	0.6774	0.7000	0.8000
2	7%	33%	0.6666	0.6666	0.7500	0.7500	0.5000	1.0000	0.6666
3	6%	27%	0.6666	0.6666	0.8000	1.0000	0.5000	0.7500	0.3333
4	16%	15%	1.0000	1.0000	0.3750	0.8000	0.8750	0.8750	1.0000
5	23%	38%	0.4545	0.4545	0.5000	0.3636	0.5000	0.5454	0.7272
Avg	–	–	0.6970	**0.7042**	0.6334	**0.7221**	0.6104	**0.7741**	**0.7055**

Table 3
Experiment 2 - Chronic Kidney Disease.

Part	Size	Pos (%)	Acc (local)	Acc Federated Averaging (Fed)	Acc Size (Fed)	Acc Accuracy (Fed)	Acc Size & acc (Fed)	Acc Cont (Fed)	Acc Inv Cont (Fed)
1	20%	73%	1.0000	0.9375	1.0000	1.0000	1.0000	0.8750	0.8583
2	20%	58%	0.9375	1.0000	0.9375	1.0000	1.0000	1.0000	1.0000
3	20%	56%	0.8125	0.8750	0.9375	0.8750	0.9375	0.9375	0.8750
4	20%	65%	0.9375	1.0000	1.0000	0.8750	1.0000	1.0000	1.0000
5	20%	60%	1.0000	1.0000	0.8750	1.0000	0.8750	1.0000	1.0000
Avg	–	–	0.9375	**0.9625**	**0.9500**	**0.9500**	**0.9625**	**0.9625**	**0.9467**
1	14%	36%	1.0000	1.0000	1.0000	0.9166	0.9000	0.9285	0.9166
2	29%	67%	0.9583	0.9583	0.9583	1.0000	0.9545	0.8947	0.9565
3	20%	77%	0.8750	0.9375	0.9285	1.0000	1.0000	0.9411	0.8571
4	13%	48%	1.0000	1.0000	0.9000	1.0000	1.0000	1.0000	1.0000
5	24%	68%	0.9500	0.9000	0.9583	0.9200	0.9545	1.0000	0.9523
Avg	–	–	0.9567	**0.9592**	0.9490	**0.9673**	**0.9618**	0.9528	0.9365
1	52%	64%	1.0000	0.9761	0.9756	0.9250	0.9767	0.9750	0.9500
2	3%	50%	1.0000	1.0000	1.0000	1.0000	0.5000	1.0000	1.0000
3	14%	76%	1.0000	1.0000	1.0000	1.0000	0.9285	1.0000	0.9285
4	5%	83%	0.7500	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
5	26%	51%	0.9523	0.9523	0.9090	0.9523	0.8333	0.9500	0.9130
Avg	–	–	0.9404	**0.9857**	**0.9769**	**0.9755**	0.8477	**0.9855**	**0.9583**
1	44%	63%	0.9428	0.9428	0.9743	0.9750	0.9736	0.9047	0.9756
2	8%	72%	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
3	6%	43%	0.8000	0.8000	1.0000	1.0000	1.0000	1.0000	1.0000
4	13%	43%	0.9090	0.9090	1.0000	1.0000	1.0000	1.0000	1.0000
5	30%	72%	1.0000	1.0000	0.8571	0.9444	1.0000	1.0000	1.0000
Avg	–	–	0.9304	0.9304	**0.9663**	**0.9839**	**0.9947**	**0.9809**	**0.9951**

Table 4
Experiment 3 - Parkinson's.

Part	Size	Pos (%)	Acc (local)	Acc Federated Averaging (Fed)	Acc Size (Fed)	Acc Accuracy (Fed)	Acc Size & acc (Fed)	Acc Cont (Fed)	Acc Inv Cont (Fed)
1	20%	65%	0.4285	0.7142	0.8571	1.0000	1.0000	1.0000	0.8571
2	20%	78%	0.7500	0.8750	1.0000	1.0000	0.8750	1.0000	1.0000
3	20%	75%	0.5000	0.8750	0.8750	0.8750	0.8750	0.8750	0.8750
4	20%	78%	0.7500	0.7500	0.8750	0.8750	0.8750	0.5000	0.7500
5	20%	79%	1.0000	1.0000	0.6250	0.8750	0.8750	0.8750	0.8750
Avg	–	–	0.6857	**0.8429**	**0.8464**	**0.9250**	**0.9000**	**0.8500**	**0.8714**
1	11%	61%	0.6000	0.8000	0.6000	0.7142	0.8000	0.8000	0.6666
2	25%	71%	0.9000	0.9000	0.9000	0.7272	0.8461	0.8000	0.7777
3	27%	87%	0.8181	0.9090	1.0000	0.8750	1.0000	1.0000	0.8000
4	10%	45%	0.5000	0.7500	0.6000	0.7500	1.0000	0.7500	0.8000
5	27%	84%	0.9090	0.9090	0.8000	0.9090	0.8181	0.8333	1.0000
Avg	–	–	0.7454	**0.8536**	**0.7800**	**0.7951**	**0.8929**	**0.8367**	**0.8089**
1	50%	78%	0.8500	0.8500	0.8181	0.8500	0.7894	0.6666	0.8500
2	6%	57%	0.6666	0.3333	1.0000	0.7500	0.6666	0.8000	0.7500
3	13%	84%	0.7142	0.8571	0.7500	1.0000	0.3333	0.7142	0.8571
4	4%	57%	1.0000	0.6666	0.6666	0.6666	0.7500	0.6666	0.3333
5	21%	79%	0.8750	0.8750	0.7777	0.6250	0.7777	0.6250	0.4285
Avg	–	–	0.8212	0.7164	0.8025	0.7783	0.6634	0.6945	0.6438
1	44%	63%	0.9428	0.9428	0.9743	0.9750	0.9736	0.9047	0.9756
2	8%	72%	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
3	6%	43%	0.8000	0.8000	1.0000	1.0000	1.0000	1.0000	1.0000
4	13%	43%	0.9090	0.9090	1.0000	1.0000	1.0000	1.0000	1.0000
5	30%	72%	1.0000	1.0000	0.8571	0.9444	1.0000	1.0000	1.0000
Avg	–	–	0.9304	0.9304	**0.9663**	**0.9839**	**0.9947**	**0.9810**	**0.9951**

This dataset was created by Max Little of the University of Oxford [16], in collaboration with the National Centre for Voice and Speech, in Denver, Colorado and is composed by a range of biomedical voice measurements from patients. Each column in the table is a particular voice measure, and each row corresponds one of 195 voice recording from these individuals.

Table 4 shows the results of the different federation processes in this dataset. In this case, the splits are more polarized, finding that with the even split and the first uneven split all aggregation methods improve the performance of the local models, while for the second uneven split, with two participants with 4% and 6% of the data, no aggregation increases the accuracy. In the last uneven split, all aggregation methods improve the local models except the Federated Averaging.

4.4. Experiment 4 - heart disease

Heart disease describes a range of conditions that affect the heart, including blood vessel diseases, such as coronary artery disease, arrhythmia and congenital heart defects among others.

Heart disease is one of the biggest causes of morbidity and mortality among the population of the world. Prediction of cardiovascular disease is regarded as one of the most important subjects in the section of clinical data analysis.

The Heart Disease dataset includes data from noninvasive test results of patients undergoing angiographies in order to study the possibility of angiographic coronary disease in them. The data collected include 14 attributes of the patients. For more information about this dataset, see [7].

In Table 5 we find the results of the different federation processes changing the aggregation method. With the even split we see that all aggregation methods improve the accuracy of the local models but for the aggregation based in a weighted average of the size of the participants and the accuracy of the local model. In the case of the first uneven split, the only aggregation methods that are able to improve the performance are the size-based and the accuracy-based. On the other hand, for the second uneven split, with two participants with 2% and 4% of the data, all aggregation methods increase the accuracy. Finally, for the last uneven split, only the Federated Averaging and the accuracy-based aggregation improve the performance of the models, resulting on the accuracy-based aggregation being the only aggregation that improves in all cases for this dataset.

5. Conclusions

Federated learning is a distributed artificial intelligence approach that can be very useful for hospitals to build collaboratively a machine learning model in a secure way, without sharing their private data. Nevertheless, the aggregation method is a decisive parameter that can change the performance of the final federated model.

In this research we have proved that the classical Federated Averaging is a reliable aggregation method that improves the performance of the local methods in 11 out of 16 cases that we have contemplated. Nevertheless, there are other aggregation

Table 5
Experiment 4 - Heart Disease.

Part	Size	Pos (%)	Acc (local)	Acc Federated Averaging (Fed)	Acc Size (Fed)	Acc Accuracy (Fed)	Acc Size & acc (Fed)	Acc Cont (Fed)	Acc Inv Cont (Fed)
1	20%	53%	0.7500	0.7500	1.0000	0.7500	0.8333	0.7500	0.9166
2	20%	64%	0.6666	0.8333	0.8333	0.8333	0.7500	0.8333	1.0000
3	20%	44%	0.5833	0.8333	1.0000	0.9166	0.5833	0.8333	0.8333
4	20%	58%	0.9166	0.8333	0.7500	0.8333	0.5833	0.9166	0.9166
5	20%	54%	0.9230	1.0000	0.7692	0.7692	0.6153	0.6923	0.6923
Avg	–	–	0.7679	**0.8500**	**0.8705**	**0.8205**	0.6731	**0.8051**	**0.8718**
1	15%	30%	0.8000	0.7000	1.0000	0.7777	0.9000	0.7142	0.6000
2	26%	63%	0.8750	0.7500	0.7500	0.7142	0.9375	0.9444	0.7647
3	18%	78%	0.9090	0.9090	1.0000	0.8333	0.7692	0.8888	0.7500
4	13%	39%	0.6250	0.8750	0.7777	0.8888	0.7000	1.0000	1.0000
5	28%	51%	0.8888	0.8333	0.9230	0.8888	0.5714	0.9047	0.8666
Avg	–	–	0.8195	0.8134	**0.8902**	**0.8206**	0.7756	0.7963	0.7963
1	50%	55%	0.6666	0.7333	0.7333	0.7500	0.7241	0.8387	0.7096
2	2%	75%	0.5000	1.0000	0.6666	1.0000	0.5000	0.6666	0.6666
3	14%	72%	0.8888	0.8888	0.8000	0.8750	0.8181	0.7272	0.7777
4	4%	71%	0.3333	0.6666	0.3333	1.0000	0.6666	0.7500	0.5000
5	30%	43%	0.6842	0.6315	0.8125	0.8235	0.5882	0.7142	0.7500
Avg	–	–	0.6146	**0.7841**	**0.6692**	**0.8897**	**0.6594**	**0.7394**	**0.6808**
1	47%	58%	0.7500	0.6875	0.8387	0.7500	0.7666	0.8928	0.8787
2	7%	38%	0.8000	0.8000	0.6000	0.8000	0.6000	0.6666	0.6000
3	10%	39%	0.4000	0.4000	0.2500	1.0000	0.2500	0.5000	0.7500
4	9%	31%	0.8750	1.0000	1.0000	0.7142	0.6666	0.8000	0.6000
5	26%	68%	0.9285	0.9285	0.8750	0.7857	0.9333	0.7777	0.8235
Avg	–	–	0.7507	**0.7632**	0.7127	**0.8100**	0.6433	0.7274	0.7305

methods with similar or even better behaviour. The contribution-based aggregation, using the difference between the losses of the global and local model, increases the accuracy in 11 out of 16 cases as well, while the size-based and the inverse contribution-based perform better in one more case. The weighted average using both the size of the participant's dataset and the accuracy of the local model increase the accuracy in only 10 cases out of 16. Finally, the weighted average using the accuracy outperforms all aggregation methods, improving the accuracy in 15 out of 16 cases, that is, in all experiments but one partition where all other methods failed to increase the performance as well.

With these results, the authors believe that an accuracy-based federated learning may perform better than the Federated Averaging classical approach. A fully connected peer-to-peer architecture has been used to show a resilient architecture against different points of failures, including the central server. As limitations of this study, using open-source datasets instead of real-world data in a study on federated learning with medical data may limit the realism, generalization, diversity, quality, and ethical considerations of the research findings.

CRediT authorship contribution statement

Jose L. Salmeron; Irina Arevalo; Antonio Ruiz-Celma: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data included in article/supp. material/referenced in article.

Acknowledgements

Prof. Salmeron research was kindly supported by the project Artificial Intelligence for Healthy Aging (Convocatoria 2021 – Misiones de I+D en Inteligencia Artificial: Inteligencia Artificial distribuida para el diagnóstico y tratamiento temprano de enfermedades con gran prevalencia en el envejecimiento, exp.: MIA.2021.M02.0007) lead by Capgemini Engineering.

References

[1] M. Abadi, A. Chu, I. Goodfellow, B. McMahan, I. Mironov, K. Talwar, L. Zhang, Deep learning with differential privacy, in: 23rd ACM Conference on Computer and Communications Security (ACM CCS), 2016, pp. 308–318, https://arxiv.org/abs/1607.00133.

[2] A. Acar, H. Aksu, S. Uluagac, M. Conti, A survey on homomorphic encryption schemes: theory and implementation, ACM Comput. Surv. 51 (04) (2017), https://doi.org/10.1145/3214303.

[3] K.M. Ahmed, A. Imteaj, M.H. Amini, Federated deep learning for heterogeneous edge computing, in: 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA), 2021, pp. 1146–1152.

[4] R.S. Antunes, C.A. da Costa, A. Küderle, I.A. Yari, B. Eskofie, Federated learning for healthcare: systematic review and architecture proposal, ACM Trans. Intell. Syst. Technol. 13 (4) (2022) 1–23, https://doi.org/10.1145/3501813.

[5] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, V. Shmatikov, How to backdoor federated learning, in: S. Chiappa, R. Calandra (Eds.), Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, in: Proceedings of Machine Learning Research, vol. 108, PMLR, 26–28 Aug 2020, pp. 2938–2948.

[6] Y. Cheng, Y. Liu, T. Chen, Q. Yang, Federated learning for privacy-preserving ai, Commun. ACM 63 (12) (Nov 2020) 33–36, https://doi.org/10.1145/3387107.

[7] R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J.J. Schmid, S. Sandhu, K.H. Guppy, S. Lee, V. Froelicher, International application of a new probability algorithm for the diagnosis of coronary artery disease, Am. J. Cardiol. 64 (5) (1989) 304–310, https://doi.org/10.1016/0002-9149(89)90524-9, https://www.sciencedirect.com/science/article/pii/0002914989905249.

[8] D. Dua, C. Graff, UCI machine learning repository, http://archive.ics.uci.edu/ml, 2017.

[9] R. Hou, S. Ai, Q. Chen, H. Yan, T. Huang, K. Chen, Similarity-based integrity protection for deep learning systems, Inf. Sci. 601 (2022) 255–267, https://doi.org/10.1016/j.ins.2022.04.003, https://www.sciencedirect.com/science/article/pii/S0020025522003279.

[10] W. Hoyos, J. Aguilar, M. Toro, Federated learning approaches for fuzzy cognitive maps to support clinical decision-making in dengue, Eng. Appl. Artif. Intell. 123 Part B (August 2023) 106371.

[11] R. Hu, Y. Guo, H. Li, Q. Pei, Y. Gong, Privacy-preserving personalized federated learning, in: ICC 2020 - 2020 IEEE International Conference on Communications (ICC), 2020, pp. 1–6.

[12] G. Kaissis, M. Makowski, D. Rückert, R. Braren, Secure, privacy-preserving and federated machine learning in medical imaging, Nat. Mach. Intell. 2 (2020) 305–311.

[13] A.M. Kermarrec, F. Taïani, Want to scale in centralized systems? Think p2p, J. Internet Serv. Appl. 6 (1) (2015) 16.

[14] J. Konecný, B. McMahan, D. Ramage, P. Richtárik, Federated optimization: distributed machine learning for on-device intelligence, arXiv:1610.02527 [abs], 2016.

[15] H. Li, Chengcheng Li, J. Wang, A. Yang, Z. Ma, Zunqian Zhang, D. Hua, Review on security of federated learning and its application in healthcare, Future Gener. Comput. Syst. 144 (July 2023) 271–290, https://doi.org/10.1016/j.future.2023.02.021.

[16] M. Little, P. Mcsharry, S. Roberts, D. Costello, I. Moroz, Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection, Biomed. Eng. Online 6 (23) (02 2007), https://doi.org/10.1186/1475-925X-6-23.

[17] Y. Liu, Y. Kang, C. Xing, T. Chen, Q. Yang, A secure federated transfer learning framework, IEEE Intell. Syst. 35 (4) (2020) 70–82, https://doi.org/10.1109/MIS.2020.2988525.

[18] B. McMahan, E. Moore, D. Ramage, B.A. y Arcas, Federated learning of deep networks using model averaging, arXiv:1602.05629 [abs], 2016.

[19] B. McMahan, D. Ramage, Google ai blog, https://ai.googleblog.com/2017/04/federated-learning-collaborative.html, Apr 2017.

[20] D.C. Nguyen, Q.V. Pham, P.N. Pathirana, M. Ding, A. Seneviratne, Z. Lin, O. Dobre, W.J. Hwang, Federated learning for smart healthcare: a survey, ACM Comput. Surv. 55 (3) (February 2022) 1–37, https://doi.org/10.1145/3501296.

[21] S.A. Rahimi, F. Légaré, G. Sharma, P. Archambault, H.T.V. Zomahoun, S. Chandavong, N. Rheault, S.T. Wong, L. Langlois, Y. Couturier, J.L. Salmeron, M.P. Gagnon, J. Légaré, Application of artificial intelligence in community-based primary health care: systematic scoping review and critical appraisal, J. Med. Internet Res. 23 (9) (2021) 1–19.

[22] J.L. Salmeron, I. Arévalo, A privacy-preserving, distributed and cooperative fcm-based learning approach for cancer research, in: R. Bello, D. Miao, R. Falcon, M. Nakata, A. Rosete, D. Ciucci (Eds.), Rough Sets, Springer International Publishing, Cham, 2020, pp. 477–487.

[23] F. Sattler, S. Wiedemann, K.R. Müller, W. Samek, Robust and communication-efficient federated learning from non-iid data, IEEE Trans. Neural Netw. Learn. Syst. 31 (9) (2020) 3400–3413, https://doi.org/10.1109/TNNLS.2019.2944481.

[24] W.N. Street, W.H. Wolberg, O.L. Mangasarian, Nuclear feature extraction for breast tumor diagnosis, in: R.S. Acharya, D.B. Goldgof (Eds.), Biomedical Image Processing and Biomedical Visualization, in: International Society for Optics and Photonics, vol. 1905, SPIE, 1993, pp. 861–870.

[25] W. Street, W. Wolberg, O. Mangasarian, Breast cancer diagnosis and prognosis via linear programming, Oper. Res. 43 (4) (Aug 1995) 570–577, https://doi.org/10.1287/opre.43.4.570.

[26] L. Su, V.K.N. Lau, Hierarchical federated learning for hybrid data partitioning across multitype sensors, IEEE Int. Things J. 8 (13) (July 2021) 10922–10939.

[27] Z. Wang, M. Song, Z. Zhang, Y. Song, Q. Wang, H. Qi, Beyond inferring class representatives: user-level privacy leakage from federated learning, in: IEEE INFOCOM 2019 - IEEE Conference on Computer Communications, 2019, pp. 2512–2520.

[28] T. Wink, Z. Nochta, An approach for peer-to-peer federated learning, in: 2021 51st Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W), 2021, pp. 150–157.

[29] T. Wink, Z. Nochta, An approach for peer-to-peer federated learning, in: 2021 51st Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W), 2021, pp. 150–157.

[30] J. Xu, B.S. Glicksberg, C. Su, P. Walker, J. Bian, F. Wang, Federated learning for healthcare informatics, J. Healthc. Inform. Res. 5 (November 2021) 1–19, https://doi.org/10.1007/s41666-020-00082-4.

[31] Q. Yang, Y. Liu, T. Chen, Y. Tong, Federated machine learning: concept and applications, ACM Trans. Intell. Syst. Technol. 10 (2) (March 2019) 1–19, https://doi.org/10.1145/3298981.

[32] Z. Zhao, C. Feng, W. Hong, J. Jiang, C. Jia, T.Q.S. Quek, M. Peng, Federated learning with non-iid data in wireless networks, IEEE Trans. Wirel. Commun. 21 (3) (March 2022) 1927–1942, https://doi.org/10.1109/TWC.2021.3108197.