Federated Learning with Accurate Model Training and Low Communication Cost in Heterogeneous Scenarios

Lucas Airam C. de Souza[1,2], **Miguel Elias M. Campista**[1], **Luís Henrique M. K. Costa**[1]

[1]Grupo de Teleinformática e Automação (GTA)
Universidade Federal do Rio de Janeiro (UFRJ)
[2]École Polytechnique, INRIA Saclay, França

Resumo. *Federated learning (FL) is a distributed approach to train machine learning models without disclosing private data from participating clients to a central server. Nevertheless, FL performance depends on the data distribution, and the training struggles to converge when clients have distinct data distributions, increasing overall training time and the final model prediction error. This work proposes two strategies to reduce the impact of data heterogeneity in FL scenarios. Firstly, we propose a hierarchical client clustering system to mitigate the convergence obstacles of federated learning in non-Independent and Identically Distributed (IID) scenarios. The results show that our system has a better classification performance than FedAVG, increasing its accuracy by approximately 16% on non-IID scenarios. Furthermore, we improve our first proposal by implementing ATHENA-FL, a federated learning system that shares knowledge among different clusters. The proposed system also uses the one-versus-all model to train one binary detector for each class in the cluster. Thus, clients can compose complex models combining multiple detectors. ATHENA-FL mitigates data heterogeneity by maintaining the clustering step before training to mitigate data heterogeneity. Our results show that ATHENA-FL correctly identifies samples, achieving up to 10.9% higher accuracy than traditional training. Finally, ATHENA-FL achieves lower training communication costs than MobileNet architecture, reducing the number of transmitted bytes between 25% and 97% across evaluated scenarios.*

1. Introduction

Problem Description: In traditional machine learning systems, model training requires client-data collection, which usually reveals private or sensitive information from the user or collection point [Liu et al. 2021]. Therefore, Federated Learning (FL) has emerged as a proposal for training machine learning models that preserve the privacy of the user without sharing local data. Federated learning (FL), proposed by Google [McMahan et al. 2017], has become popular among researchers and industry due to the possibility of training machine learning models while preserving users' data privacy.

Training in federated learning replaces data sharing with model parameter sharing. In Federated Averaging (FedAVG), the most widely used algorithm for model parameters' aggregation in FL, clients train the model locally for a few epochs and send the results to the aggregation server, which combines the individual trained models into a single global model. Thus, clients' training samples remain stored locally, preserving data and user privacy. Nevertheless, another challenge persists when deploying FL on a large scale: clients may have heterogeneous

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001, CNPq, FAPERJ (E-26/203.211/2017, E-26/202.932/2017, E-26/202.689/2018 e E-26/200.892/2021) e FAPESP (15/24494-8, 18/23292-0, 15/24485-9 e 14/50937-1.)

data generated from non-independent and identically distributed (non-IID) distributions, leading to convergence difficulty and suboptimal performance [Ma et al. 2022, Zhao et al. 2018].

Motivation: A proposal that reduces the effects of data heterogeneity during model training and allows the generalization of the classifier to samples originating from other data distributions becomes necessary. It can be achieved by clustering clients comparing data similarity, which allows models to be trained on IID data and quickly converge with high final classification performance.

Objectives: Our purpose is to design a system to increase the performance of federated learning when clients have non-IID data distributions. The first goal is to tackle the non-IID problem by clustering clients according to the neural network weights. This procedure generates clusters in which clients hold similar data distributions. The second challenge is to create an intra-cluster information sharing scheme. Since clients might need to classify data generated with different distributions, information sharing is a key factor to obtain generic models.

Contributions: This work proposes a hierarchical client clustering system to increase the efficiency of federated learning in scenarios where clients have non-IID datasets. Clients, also called nodes, are divided into clusters where the data are similar. The proposal uses clients' last-layer neural network weights as input to perform clustering. The last-layer neural network weights maintain statistical relationships with the clients' private data without revealing them [Wang et al. 2020]. Thus, we maintain the same privacy model as FedAVG. Firstly, each cluster trains a personalized model with independent hyperparameters and parameters, allowing high classification performance on specific tasks. We also investigate how to tune the clustering algorithms' hyperparameters to best fit our scenario. Then, we extend the first proposal by providing a mechanism to combine models created on different clusters, which we call ATHENA-FL (Avoiding sTatistical HEterogeiNety with one-*versus*-All in Federated Learning), a system that creates models with the one-*versus*-all (OvA) technique to enable model sharing between groups. The one-*versus*-all model uses independently trained binary classifiers, which estimate the probability that a sample belongs to the class identified by a detector. After training, the detectors are combined for sample classification. Each detector estimates the probability that the sample belongs to its class, and the classifier labels the sample from the detector that generates the highest probability. Thus, the OvA method is used for efficient model sharing between groups to create a generic model for classifying data from different groups.

We analyze different clustering models and show the advantage of adopting the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm for client clustering. DBSCAN can identify homogeneous and heterogeneous data distributions from the client neural network weights and outperforms the Ordering Points to Identify the Clustering Structure (OPTICS) and K-Means algorithms. Furthermore, in the worst case, our results show that the proposed approach performs equally to traditional federated learning. For non-IID datasets, the FedAVG algorithm converges with low accuracy after 100 global epochs. Nonetheless, the current proposal achieves better results across all clusters in 10 global epochs, providing higher classification performance. Then, we evaluate the ATHENA-FL proposal according to the accuracy evolution of the detectors. Also, the evaluation compares the classification performance of the global model and communication requirements for training the models. The accuracy of the models depends on the diversity of the samples used for training the detectors, with the accuracy of the OvA model up to 10.9% higher than the MobileNet architecture. At the same time, the amount of bytes transmitted over all training epochs is reduced by up to 97.37% using the one-*versus*-all model instead of MobileNet. Thus, the system provides an effective way to train models that maintain the privacy of client data in scenarios with heterogeneous distributions.

We summarize our contributions as follows:

- **High classification performance**: We propose a system that achieves high accuracy results under non-IID data distributions. Our proposal obtains these results by assigning clients to a cluster where the data is more homogeneous.
- **Low-communication requirement**: The proposal uses clients' last-layer neural network weights as the input to perform client clustering. Thus, we avoid the communication overhead of recursive clustering proposals. Also, we reduce the amount of data used to cluster clients compared to a proposal that sends a whole model.
- **Privacy-preserving**: Our proposal holds the same privacy model and assumptions as FedAVG, given that the system shares only the model weights.
- **Robust-classifier**: The OvA model is used to combine the classifiers trained among different clusters, which offers a classifier that is able to identify samples generated in different clusters.
- **Prototype**: We implement a proof-of-concept of our proposed system. We also present a practical evaluation of FedAVG under non-IID data distributions and compare it with our proposal.

2. Related Work

Personalized Models in Federated Learning: Federated learning by Transformer Personalization (FedTP) is a framework for reducing data heterogeneity by transforming datasets and personalizing models [Li et al. 2023]. FedTP objective is to define the basis of a transformation. In this transformation, the client data is similar, and they can personalize their models in some layers, overcoming the convergence problems. Federated Learning Early Exit of inference (FLEE) is a hierarchical-federated learning framework that splits the model into three different geolocations [Zhong et al. 2022]. The model's division between the cloud, edge, and end device allows using the method of early exit inference from neural networks. Furthermore, FLEE's hierarchical division of the training reduces the impact caused by non-IID data distributions on the model convergence, as the authors assume that the data have higher similarity according to the geographic distance of the clients.

Zhu *et al.* propose using one-*versus*-all classification scheme to mitigate the impact of heterogeneous data in training federated learning models [Zhu et al. 2021]. Their Federated OvA (FedOVA) algorithm uses models for binary classification and selects clients having samples of the target class to perform the training. Nevertheless, their proposal lacks a study on the impact of the model creation and an adequate protocol for detector training. ATHENA-FL clusters clients according to data distributions, before training the OvA model, thus reducing detector training time.

Clustered Federated Learning: Clustered federated learning is a subfield of model personalization in federated learning research. This subfield focuses on creating strategies to arrange FL clients into groups, reducing data heterogeneity. The proposals are mostly differentiated by clustering strategy, metrics, and training. Communication-Efficient Federated Learning (CEFL) is an FL-based framework for medical-data model training [Chu et al. 2022]. The authors determine the similarity between clients by calculating the Euclidean distance of the clients' neural network weights. Based on similarity, the system clusters clients using the Louvain method [Blondel et al. 2008]. In each group, the client with the highest sum of similarity is the leader. The leader performs federated training of the model's first layers with leaders from other groups. The model's final layers are trained individually in each group.

Stochastic Clustered Federated Learning (StoCFL) [Zeng et al. 2023] is a federated learning framework that clusters clients according to the cosine similarity. The proposal introduces two models: the global model and the cluster model. The global model maintains information from all clusters, while clients only participate in model training inside the cluster they are involved. Meanwhile, there is a high cost to the system's clients, who must simultaneously train the two models. Iterative Federated Clustering Algorithm (IFCA) [Ghosh et al. 2020] is a proposal for clustering clients to personalize their models. In the proposal, clients are responsible for choosing their clusters. In addition, the authors propose the use of multi-task learning, which consists of sharing some neural network weights for clients who have data distributions with intersections but are in different clusters. Nevertheless, the downside of delegating the process of identifying groups to the clients is that the environment becomes susceptible to malicious behavior and requires more computational costs from the clients' devices. Another disadvantage of this proposal is to assume that the number of groups is known a priori, which may be unfeasible and can either overestimate or underestimate the number of existing clusters.

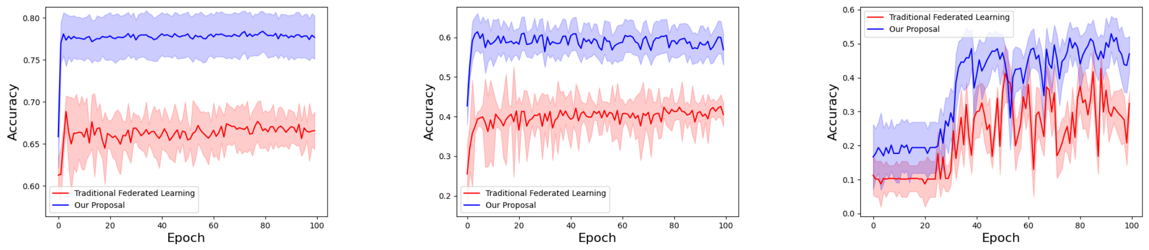
Clustered Federated Learning (CFL) [Sattler et al. 2020] recursively partitions federated learning clients into more homogeneous groups. This procedure mitigates the problems generated by non-IID distributions. Partitioning occurs whenever the loss gradient vector exceeds a pre-established distance threshold. Nonetheless, clients' recursive partition leads to computational overhead on the aggregation server because it runs the procedure at each global training epoch. Flexible Clustered Federated Learning (FlexCFL) [Duan et al. 2022] is a framework that considers clients' data distribution time shifts. The grouping strategy is static, which avoids rescheduling clients for each epoch, as [Sattler et al. 2020] and [Ghosh et al. 2020] do. Also, the framework uses only a subset of clients to determine the number of clusters in the system and calculate the Decomposed Cosine Similarity (DCS). The remaining clients are clustered after the decision about the number of clusters. Before each round, the authors execute a strategy to detect if the clients remain with the same distribution or if it has changed. When the distance exceeds a predefined threshold, the client needs to perform the clustering step again. Nevertheless, the proposal requires higher computational resources than our proposal to detect the data shift.

3. Results

The experiments were executed on an Intel Xeon CPU E5-2650 2.00 GHz server with 32 processing cores and 504 GB of RAM. We show experimental results of the models' evaluation, with the average accuracy obtained among all the clients and a 95% confidence interval. We compare our proposal with FedAVG because other approaches use different models or datasets. Thus, FedAVG establishes a baseline comparison with the state-of-the-art algorithms.

3.1. Clustering Evaluation

In the first experiment, the data distributions on the clients are non-IID, with only two classes present in their datasets. The goal of this experiment is to compare the performance of the model generated through traditional federated learning with the one created by the current proposal, in which the data distribution is non-IID. Thus, the 6,000 samples of a class are split among five clients to create balanced datasets. Fig. 1(a) displays the experimental mean result among all clusters, illustrated in the figure as the blue line, and the traditional result, represented by the red line. Traditional federated learning is affected by the level of data heterogeneity, while the current proposal allows for high classification performance. The experiment also analyzes the individual behavior of the clusters' models. The results show that the first cluster has an accuracy of (56 ± 3)%, while the second has an accuracy of (63 ± 3)%. Although there is a cluster with higher performance, even the low performance of the first cluster is higher than that obtained when using the traditional proposal. Thus, it can be concluded that the proposal successfully mitigated the effects of heterogeneity in the scenario with five classes per client.



(a) 2 classes per client.

(b) 5 classes per client.

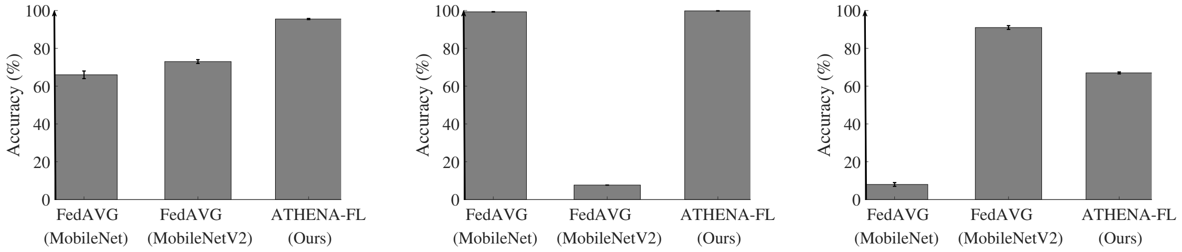
(c) LDA with $\alpha - 0.5$.

Figure 1: Models' accuracy under non-IID data distributions with CIFAR-10 dataset.

The second experiment evaluates the performance of the proposal for non-IID datasets with 5 classes per client. The results shown in Fig. 1(b) show similar behavior to the first experiment for both cases, but with a worse final performance. This fact can be explained by the difficulty of the problem increased in this case.

The third experiment evaluates the clustering approach in a non-IID scenario where the clients' datasets are not limited to samples of a subset of classes, thus creating a more realistic scenario. We build the datasets of this experiment through the Label-based Dirichlet Partition (LDA), which generates the data based on the number of clients and non-IID degree. Figure 1(c) shows the results of Experiment V. Our approach increases by more than 14% on average the accuracy compared with traditional federated learning.

3.2. ATHENA-FL Evaluation



(a) CIFAR-10 dataset.

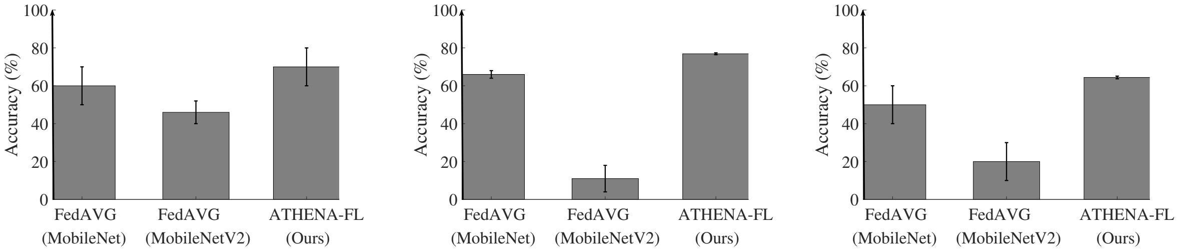
(b) MNIST dataset.

(c) FMNIST dataset.

Figure 2: Models' accuracy evaluation in the IID setting.

IID data The architecture MobileNetV2 has an accuracy of $(73 \pm 1)\%$ and the MobileNet has (66 ± 2) in the CIFAR-10 dataset, as exhibited in Figure 2(a). In the MNIST dataset, the performance is better, with $(99.3 \pm 0.1)\%$ and $(99.8 \pm 0.1)\%$ for MobileNet and ATHENA-FL, as shown in Figure 2(b). However, MobileNetV2 is too complex for those data and has an overfitting problem, leading to only (7.69 ± 0.01) accuracy. Finally, the FMNIST shows that MobileNetV2 has the best performance in the IID setting, with (91 ± 1) against only (8 ± 1) for the MobileNet and (67.0 ± 0.5) for ATHENA-FL, as exhibited in Figure 2(c).

Thus, the experiment shows that under IID settings, ATHENA-FL has competitive results, but in some scenarios, the detectors might need to be well-adjusted to have a better performance. The variation in performance between the datasets is due to the difficulty of the problem presented by each one. MNIST has simpler grayscale images, which are simpler for classification. Therefore, the detectors have higher accuracy and less variance in this dataset than in CIFAR-10, which has color images with more elements. Finally, the shapes of clothes in the FMNIST dataset are difficult to distinguish. Thus, we need more complex models to differentiate them. This behavior is also observed in the other evaluated scenarios.



(a) CIFAR-10 dataset.

(b) MNIST dataset.

(c) FMNIST dataset.

Figure 3: Models' accuracy evaluation with Non-IID distribution of 5 classes per client.

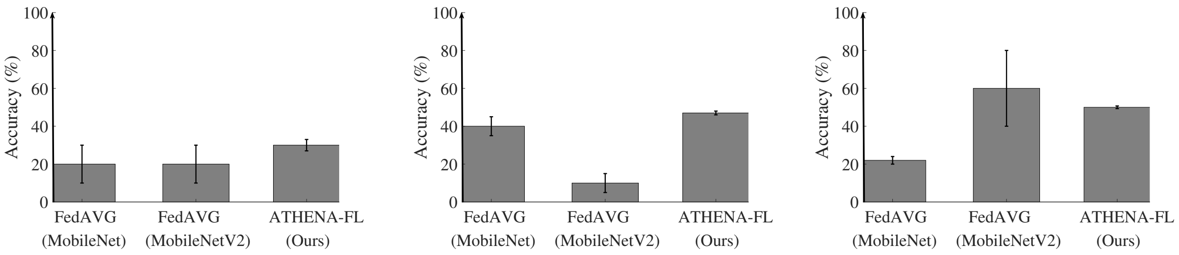
Non-IID data Scenarios with non-IID data consider distributions of data where clients own only a subset of the classes of the dataset. In the first non-IID case, clients have samples from five distinct classes. Figure 3(b), 3(c), and 3 show the final performance of the one-*versus*-all model, MobileNet and MobileNetV2. ATHENA-FL provides the best accuracy results for all datasets in this setting, having with the CIFAR-10 dataset 10% higher accuracy compared to the MobileNet, which is the second-best model, and 10.9% higher accuracy in the MNIST dataset. Thus, the experiment demonstrates that ATHENA-FL has the potential to increase the classification accuracy up to 10.9% under the Non-IID setting compared to the MobileNet model trained purely with FedAVG.

The last scenario considers a non-IID data distribution with two classes per client. For the CIFAR-10 dataset, the observed accuracy is $(30 \pm 3)\%$ combining the detectors, while the MobileNet and MobileNetV2 architectures have a final accuracy of $(20 \pm 10)\%$ in this configuration, shown in Figure 4(b). The accuracy of ATHENA-FL was $(47 \pm 1)\%$ for the MNIST dataset, while the deep models achieved an accuracy of $(40 \pm 10)\%$ and $(10 \pm 5)\%$ for MobileNet and MobileNetV2 respectively, exhibited in Figure 4(c). Lastly, MobileNet has $(22 \pm 2)\%$, MobileNetV2 has $(60 \pm 20)\%$, and ATHENA-FL has $(50.0 \pm 0.7)\%$ accuracy in the FMNIST dataset, in Figure 4.

The results show that when the detectors are trained on datasets with more classes, the final classification performance is better since they can learn more patterns of the whole data distribution. This behavior can be explained by the higher variance of data from classes that are not of the detector's class. The greater variance of data in other classes allows the detector to identify more relevant features in the data of interest, instead of just differentiating specific image features that are not representative of the problem. For instance, the detectors trained in the MNIST dataset with only classes 2 and 3 have trouble differentiating 5 and 8 which have similar shapes. Nonetheless, we see that in MNIST and CIFAR-10, ATHENA-FL was able to reach up to 7% and 10% accuracy compared to the best deeper model.

Communication Evaluation The objective of this experiment is to evaluate the cost of communication between the clients and the aggregation server while training the one-*versus*-all models and a deep neural network to classify multiple classes. The communication evaluation considers the total number of bytes transmitted on average to perform model training.

After estimating the number of epochs necessary for the models to converge, it is possible to compare the communication cost. The results show how many bytes each neural network



(a) CIFAR-10 dataset.

(b) MNIST dataset.

(c) FMNIST dataset.

Figure 4: Models' accuracy evaluation with Non-IID distribution of 2 classes per client.

model needs to transmit during the training. In the IID setting, the detectors converge in a few epochs, reducing over 75% of bytes transmitted. A second highlight is the data used to train and test the models. ATHENA-FL converges faster than MobileNet in MNIST and FMNIST, which have grayscale images and, therefore, use a single channel to represent the pixels. We observe that in all scenarios, ATHENA-FL reduces the total amount of bytes transmitted during the training execution. In the worst-case scenario, our approach saves at least 25% of the bytes transmitted. For the best case scenario, we can reduce by approximately 97% the communication requirements for training the model.

Table 1: Communication requirements, Total Data Transmitted (TDT), to train the models until convergence in different datasets and sample distribution settings.

		IID		**Non-IID 5**		**Non-IID 2**	
Dataset	**Model**	$\mathbb{E}[e]$ **(#)**	**TDT (MB)**	$\mathbb{E}[e]$ **(#)**	**TDT (MB)**	$\mathbb{E}[e]$ (#)	**TDT (MB)**
CIFAR-10	ATHENA-FL	7	37.635	12	64.517	14	75.270
	FedAVG	30	378.106	23	289.881	8	100.828
	Economy (%)	-	**90.05**	-	**77.74**	-	**25.35**
MNIST	ATHENA-FL	4	21.506	24	129.034	4	21.506
	FedAVG	7	88.225	58	731.005	65	819.230
	Economy (%)	-	**75.62**	-	**82.35**	-	**97.37**
FMNIST	ATHENA-FL	5	26.88	81	435.49	21	112.91
	FedAVG	44	554.56	151	1903.10	80	1008.30
	Economy (%)	-	**95.15**	-	**77.12**	-	**88.8**

4. List of Publications

This master thesis is a set of the following publications, we also have submitted two papers for an international conference and a journal that are currently under review:

- de Souza, L. A. C., Camilo, G. F., Sammarco, M., Campista, M. E. M., Costa, L. H. M. - "Aprendizado Federado com Agrupamento Hierárquico de Clientes para Aumento da Acurácia". In Anais do XL Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (pp. 545-558). (2022, May). SBC.
- de Souza, L. A. C., Camilo, G. F., Rebello, G. A. F., Sammarco, M., Campista, M. E. M., Costa, L. H. M. - "ATHENA-FL: Evitando a Heterogeneidade Estatística através do Um-contra-Todos no Aprendizado Federado". In Anais do VII Workshop de Computação Urbana (pp. 40-53). (2023, May). SBC. **Honarable Mention**.

References

Blondel, V. D. et al. (2008). Fast Unfolding of Communities in Large Networks. *Journal of Statistical Mechanics: Theory and Experiment*, pages 1–12.

Chu, D., Jaafar, W., and Yanikomeroglu, H. (2022). On the Design of Communication-Efficient Federated Learning for Health Monitoring. *IEEE GLOBECOM*, pages 1–6.

Duan, M. et al. (2022). Flexible Clustered Federated Learning for Client-Level Data Distribution Shift. *Transactions on Parallel and Distributed Systems*, 33(11):2661–2674.

Ghosh, A., Chung, J., Yin, D., and Ramchandran, K. (2020). An Efficient Framework for Clustered Federated Learning. *arXiv preprint arXiv:2006.04088*.

Li, H., Cai, Z., Wang, J., Tang, J., Ding, W., Lin, C.-T., and Shi, Y. (2023). FedTP: Federated Learning by Transformer Personalization. *IEEE Transactions on Neural Networks and Learning Systems*.

Liu, B. et al. (2021). When Machine Learning Meets Privacy: A Survey and Outlook. *Computing Surveys (CSUR)*, 54(2):1–36.

Ma, X., Zhu, J., Lin, Z., Chen, S., and Qin, Y. (2022). A State-of-the-Art Survey on Solving Non-IID Data in Federated Learning. *Future Generation Computer Systems*, 135:244–258.

McMahan, B. et al. (2017). Communication-efficient Learning of Deep Networks from Decentralized Data. *Artificial Intelligence and Statistics*, pages 1273–1282.

Sattler, F., Müller, K.-R., and Samek, W. (2020). Clustered Federated Learning: Model-Agnostic Distributed Multitask Optimization under Privacy Constraints. *IEEE Transactions on Neural Networks and Learning Systems*.

Wang, H., Kaplan, Z., Niu, D., and Li, B. (2020). Optimizing Federated Learning on Non-IID Data with Reinforcement Learning. In *IEEE INFOCOM*, pages 1698–1707.

Zeng, D., Hu, X., Liu, S., Yu, Y., Wang, Q., and Xu, Z. (2023). Stochastic Clustered Federated Learning. *arXiv preprint arXiv:2303.00897*.

Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., and Chandra, V. (2018). Federated Learning with Non-IID Data. *arXiv preprint arXiv:1806.00582*.

Zhong, Z. et al. (2022). FLEE: A Hierarchical Federated Learning Framework for Distributed Deep Neural Network over Cloud, Edge and End Device. *ACM TIST*, pages 1–24.

Zhu, Y., Markos, C., Zhao, R., Zheng, Y., and James, J. (2021). FedOVA: One-vs-All Training Method for Federated Learning with Non-IID Data. In *IEEE IJCNN*, pages 1–7.