Wireless Federated Learning (WFL) for 6G Networks - Part I: Research Challenges and Future Trends

Pavlos S. Bouzinis, *Student Member, IEEE*, Panagiotis D. Diamantoulakis, *Senior Member, IEEE*, and George K. Karagiannidis, *Fellow, IEEE*

Abstract—Conventional machine learning techniques are conducted in a centralized manner. Recently, the massive volume of generated wireless data, the privacy concerns and the increasing computing capabilities of wireless end-devices have led to the emergence of a promising decentralized solution, termed as *Wireless Federated Learning (WFL)*. In this first of the two parts paper, we present the application of WFL in the sixth generation of wireless networks (6G), which is envisioned to be an integrated communication and computing platform. After analyzing the key concepts of WFL, we discuss the core challenges of WFL imposed by the wireless (or mobile communication) environment. Finally, we shed light to the future directions of WFL, aiming to compose a constructive integration of FL into the future wireless networks.

Index Terms—Wireless Federated learning, Distributed Artificial Intelligence, 6G Networks

I. Introduction

THE sixth generation of wireless networks (6G), which is envisioned to be the evolution of wireless networks from "connected things" to "intelligent things", can be seen as an integrated communication and computing platform, with the capability to serve a vast amount of heterogeneous internet-of-things (IoT) applications, e.g, autonomous vehicles, augmented and virtual reality, smart grids, intelligent industry, smart farming, etc. Thus, 6G is expected to heavily rely on ubiquitous artificial intelligence services and gradually overwhelm the capabilities of the fifth generation (5G) of wireless networks. To this direction, the main pillar is the twofold use of artificial intelligence, both as a means to efficiently orchestrate the wireless networks and as the core of the operation of most applications. This separation steers the research in different directions, namely: i) the optimization of wireless networks performance by using machine learning (ML) techniques and ii) the enhancement of data-driven applications that are based on ML, by the joint design and optimization of communication and computing networks [1]. This work mainly focuses on the second direction, although the use of techniques from the first direction is also considered.

Nowadays, the standard ML techniques are based on a centralized concept, where the data are uploaded and processed on a single entity, e.g., a central server. However, the strict latency requirements and the data privacy assurance, renders the centralized configurations impractical for forthcoming 6G-enabled applications, such as smart grids, autonomous vehicles, and augmented reality. Hence, the combination of the aforementioned limitations with the growing computational

P. S. Bouzinis, P. Diamantoulakis, and G. Karagiannidis are with Wireless Communication and Information Processing Group (WCIP), Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Greece, E-mails: {mpouzinis, padiamian, geokarag}@auth.gr

capabilities of devices, paves the way towards implementing distributed frameworks for the construction of learning models. In the decentralized solutions, devices collaboratively train a model by leveraging their local computational resources. Among the decentralized approaches, federated learning (FL) has been proposed as a promising solution for providing distributed artificial intelligence services, while protecting the data privacy and meeting the low-latency demands [2], [3], which are key requirements of 6G networks.

The prominent feature of FL is the retention of the training dataset in the source of generation, i.e., the device. More specifically, each learner performs the model training through its local dataset individually, and forwards only the training parameters to the central server, instead of sending the overall raw data. In this manner, the central unit has no explicit access to privacy-sensitive data. Following that, the server aggregates the received parameters, updates the global model, and finally broadcasts it to the learners, while the considered process is repeated until the convergence of the global model. The principal advantages that FL is capable of providing, in accordance with the multilateral demands of 6G networks, are discussed below.

- *Privacy*: As mentioned previously, users do not share their raw data with the server or any of the residual participants. Therefore, the privacy-preserving mechanism constitutes an inherent characteristic of FL.
- *Very low Latency*: Since no raw data is sent to the cloud, the amount of information transmitted into the network is reduced, which also decreases the communication cost. Furthermore, decisions and model training can be executed locally on the end devices, instead of being sent to the server, leading to decreased latency.
- *System Heterogeneity*: The devices participating in the learning process, might present heterogeneity in terms of computational, communication resources, and data heterogeneity, which deals with non-independent and identical distribution (non-i.i.d.) of data among users. FL has the potential to tackle with the former issues.

Although recent works have successfully discussed the challenges imposed by FL and its future trends [4], [5], our contribution lies in elaborating on the wireless aspect of FL, while also shedding light on the future research directions of wireless federated learning (WFL), which are highly connected with the 6G vision. In general, the performance of WFL heavily depends on the performance of the utilized wireless communication network, while improving its performance
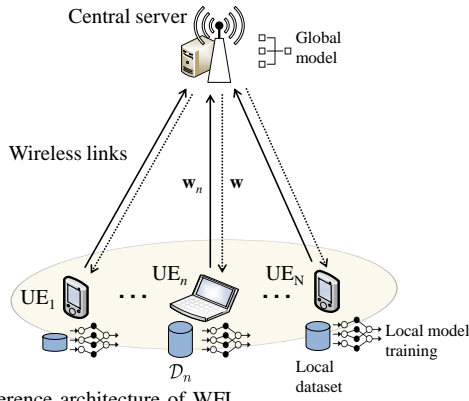
Fig. 1. Reference architecture of WFL.

requires the joint optimization of the communication and computing resources.

Next, the general WFL reference architecture is presented, which is the basis of the considered core WFL applications, as well as the analysis presented in the second part of this work.

A. Wireless federated learning reference architecture

FL refers to training a shared model in a distributed manner, by exploiting the collected data of the mobile devices without those being intervened by the server. Hence, each device contributes to the construction of the model by performing local training on its dataset, while the server's role is to aggregate, update, and redistribute the updated model back to the users. As a result, each user benefits from the local datasets of the residual participants, with the aid of a central server, while the data privacy is preserved.

As shown in Fig. 1, we consider a WFL network, which consists of N users indexed as $n \in \mathcal{N} = \{1, 2, ..., N\}$ and a BS-server. Each user n has a local dataset $\mathcal{D}_n = \{\boldsymbol{x}_{n,k}, y_{n,k}\}_{k=1}^{D_n}$, where $D_n = |\mathcal{D}_n|$ are the data samples, $\boldsymbol{x}_{n,k} \in \mathbb{R}^d$ is the k-th input data vector of user n, while $y_{n,k}$ is the corresponding output. The whole dataset is denoted as $\mathcal{D} = \bigcup_{n \in \mathcal{N}} \mathcal{D}_n$, while the size of all training data is $D = \sum_{n=1}^{N} D_n$. Also, the local loss function on the data set D_n, defined as

$$F_n(\boldsymbol{w}) \triangleq \frac{1}{D_n} \sum_{k \in \mathcal{D}_n} f(\boldsymbol{w}, \boldsymbol{x}_{n,k}, y_{n,k}), \quad \forall n \in \mathcal{N}, \quad (1)$$

where $f(\boldsymbol{w}, \boldsymbol{x}_{n,k}, y_{n,k})$ captures the error of the model parameter $\boldsymbol{w}$ for the input-output pair $\{\boldsymbol{x}_{n,k}, y_{n,k}\}$. The training process aims to find the global model parameter $\boldsymbol{w}$ which minimizes the loss function on the whole data set, which is given by $J(\boldsymbol{w}) = \frac{1}{D} \sum_{n=1}^{N} D_n F_n(\boldsymbol{w})$.

The whole training process is divided in an arbitrary number of communication rounds, denoted by i. Thus, during the first round the server initializes $\boldsymbol{w}^0$, while the i-th round is described by the following steps [3]:

i) The BS broadcasts wirelessly the global parameter $\boldsymbol{w}^i$ to all users during the considered round.

ii) After receiving the global model parameter, each user $n \in \mathcal{N}$, train its local model by applying a few steps of the gradient descent method, i.e., $\boldsymbol{w}_n^{i+1} = \boldsymbol{w}^i - \eta \nabla F_n(\boldsymbol{w}^i)$, where η is the learning rate, and then uploads the local parameter $\boldsymbol{w}_n^{i+1}$ to the server. It is noted that alternative

methods could be also employed for the local training, such as stochastic gradient descent (SGD).

iii) After receiving all the local parameters, the server aggregates them, in order to update the global model parameter, by applying $\boldsymbol{w}^{i+1} = \frac{1}{D} \sum_{n \in \mathcal{N}} D_n \boldsymbol{w}_n^{i+1}$.

The above process is repeated until a required global accuracy is achieved.

B. Applications of wireless federated learning

The range of WFL applications is quite large and not fully explored yet. Next, three core applications in the era of 6G are discussed, emphasizing on their particularities and requirements.

1) Smart grids: Smart grids can be seen as the superposition of electricity and communication networks, which enables the two-way flow of power and data, facilitating the active participation of all users in the energy management, the precise prediction of energy consumption, the avoidance of security risks, the self-healing procedure, etc. This approach leads to the generation and the requirement of processing of an enormous amount of data, which might be difficult or even impossible to be stored and processed centrally [6]. Also, despite the important benefits of the electricity networks intellification, the exchange of information between different entities and the processing of data at the cloud, exposes the smart grids to potential security and privacy risks. To overcome these challenges, the iterative local processing of data by the smart meters and aggregators at the edge and the global exploitation of the corresponding output in a collaborative manner via WFL can be particularly useful [7].

2) Unmanned mobility: Autonomous vehicles are an emerging application which is envisioned to be realized with the aid of 6G wireless networks. In order to support coordination among vehicles and satisfy the requirements of unmanned vehicular networks, machine learning-based techniques constitute a significant tool. Relevant applications are the collaborative autonomous driving, collision avoidance systems, visual object detection and traffic congestion control. Usually, the model training is performed at a central cloud in an off-line manner. However, such approaches cannot adapt to the dynamic system changes. Therefore, WFL could alleviate this burden, as a highly adaptive technique which monitors the environmental changes in real-time. Furthermore, each vehicle could benefit from the rest vehicles' observations, in a collaborative manner, leading to increased environmental knowledge [8]. At last, FL has the potential to reduce the data traffic, which is vital for latency-critical applications such as autonomous vehicles.

3) Augmented reality (AR): In the past years, AR technology has received significant attention. AR provides an interactive experience to the user, by combining virtual contents with the real world [9]. Traditionally, AR models are trained in a centralized manner. However, the latency-sensitive AR applications impose new challenges, while the centralized machine learning approach becomes non applicable. Therefore, FL is capable of providing low-latency for object detection tasks and classification problems. Also, in [9], the FL concept has been combined with mobile edge computing (MEC), in order to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LCOMM.2021.3121071, IEEE Communications Letters

3

exploit the computing capabilities of edge nodes and reduce the computational power consumption at the end device during the WFL process.

II. CHALLENGES

A. Resource allocation and participants selection

In a wireless FL environment, one of the major issues that needs to be addressed is the management of the available computation and wireless resources that users share. Recent architectures of machine learning models consist of million weight parameters, while their transmission to the server could cause significant delay. Thus, optimal resource allocation strategies can lead to decreased latency per round and subsequently to fast convergence of the global FL model. Firstly, the scarce spectrum resources ought to be efficiently allocated among users. Furthermore, since many devices are energy-limited, the power control and the computation energy consumption of the devices have to also be considered. Therefore, in order to meet the strict latency, energy and training efficiency requirements, the bandwidth allocation, the transmission power control and the devices' central processing unit (CPU) frequency clock speed adjustment issues, need to be jointly orchestrated. It is worth noting, that the resource allocation problem is highly related with the multiple access protocol selection, which we will discuss later on this work. For instance, authors in [10] minimized the total energy consumption of all users during an FL task, under latency constraints, by optimally managing the available computation and communication resources.

Besides resource management, the number of participating users in the FL task has to be tactfully selected. During a communication round, the server waits until all participants terminate the update and upload of their locally trained models. Therefore, the delay of a round is determined by the slowest device. As a result, devices with limited computation capabilities or poor wireless channel conditions, termed as stragglers, are responsible for the occurrence of long delays and can negatively affect the convergence speed. In addition to this, the limited number of resource blocks may not be adequate to support an increased number of devices. Moreover, active devices may be obligated to drop out the training process in an arbitrary time instant, due to connectivity issues related with bad channel conditions or energy-intensive tasks. Furthermore, the non-i.i.d. level of the overall dataset among users, impacts the number of selected clients. As a consequence, only a subset of the eager-to-participate users may be scheduled for participation. Thus, device scheduling policies become crucial for satisfying the underlying latency constraints, accelerating the model convergence and improving the model performance. Finally, to circumvent those challenges, asynchronous communication schemes may be considered, where the server does not necessarily wait for all participants to finish the parameter transmission [11].

B. Tradeoff between latency per round and number of rounds

In a FL task, one of the objective goals is to minimize the convergence time, in order to achieve a certain global accuracy. The convergence time is a function of the total number of rounds and the latency per round, which is subject to both computation and communication delay. As a matter of fact, during the local model training process, the number of local iterations that each device performs has to be wisely selected. Increased number of local iterations may lead to decreased number of required rounds, in the expense of energy consumption and larger latency per round. On the other hound, the execution of few local iterations are energy-saving and achieve smaller latency per round, however an increased number of total communication rounds may be enforced, in order to achieve the required global accuracy. Moreover, the considered tradeoff is also present during the participant selection procedure. By scheduling a large amount of users for participation, an increased latency per round is expected owing this to the straggler effect and the reduced bandwidth allocation. On the contrary, by selecting a few users to participate, it is more likely that the latency per round is decreased. However, in such case, the convergence speed and the model accuracy might be negatively affected, since a limited number of scheduled users contributes with a smaller dataset throughout the training process.

C. Tradeoff between model performance and convergence speed

Apart from convergence rate, global model accuracy is of paramount importance, since it is the primary goal of the training process. The inherent unreliability of wireless links can impact the quality of the WFL performance. Motivated by such considerations, authors in [12] investigated the effect of packet transmission errors, aiming to improve the FL performance, by jointly optimizing the computation and radio resources, as well as the user selection. Furthermore, in [13], the global loss function of the FL model is minimized subject to convergence time constraints. Emphasis is given to the bandwidth allocation and the user scheduling policy. Moreover, the non-i.d.d. level among the local datasets has be shown to significantly impact the user selection decisions, for a timely-efficient model performance improvement. Finally, the number of local updates that each device executes can affect the global model performance. By performing a few local iterations, a decreased global performance may occur. Reversely, the local over-optimization, i.e., large number of local updates, could lead to divergence and deterioration of the global model accuracy [3]. As a matter of fact, a balanced number of local updates should be selected, while the tradeoff between convergence speed and model accuracy should be also considered.

D. Privacy and security

Guaranteeing the privacy of the local datasets is a fundamental driving factor for implementing WFL, which can also materialize the 6G vision, as an enhanced privacy preserving architecture. Although in WFL the participants do not share their raw data, sensitive information may still be revealed to malicious third-partys or the central server e.g., with gradient leakage attacks to steal the devices' local data. The privacy in FL systems can be classified into the following two categories: global privacy and local privacy. Global privacy requires that no third party can access the global model during each

communication round, while local privacy requires that the updates of the model are also private to the server. Differential Privacy (DP) techniques [14], have been proposed to protect gradient information, while they can be used for ensuring both global and local privacy. DP is based on the addition of artificial noise in the training parameters by using a differential privacy-preserving randomized mechanism. Although DP can enhance privacy, it may sacrifice the model performance.

Moreover, FL faces security issues, such as data and model poisoning attacks. Malicious participant can send incorrect information or false models in order to undermine the training efficiency and degrade the global model performance. Thus, protection mechanisms should be constructed, which aim to detect abnormal user behavior and finally prevent malicious users from participating into the training process.

III. FUTURE RESEARCH DIRECTIONS

A. Advanced multiple access for WFL

During the model transmission phase in each communication round, all devices upload their locally trained results to the central server. Thus, the efficient integration of WFL in 6G depends on the utilized multiple access scheme. In the recent literature, mostly orthogonal multiple access (OMA) schemes are selected for the considered uplink transmission, such as frequency division multiple access (FDMA) and time division multiple access (TDMA). In the former scheme, each user occupies a sub-channel from the available bandwidth, while in TDMA, users are transmitting their messages in non-overlapping time slots by utilizing the whole available bandwidth. Moreover, *over-the-air* FL has been proposed as an efficient method for fast model aggregation, which exploits the superposition property of the wireless multiple access channel for computing the global model, resulting in reduced latency. However, this scheme is based on an analog transmission, which may be a limiting factor for modern wireless systems.

In the last years, non-orthogonal multiple access (NOMA) has drawn considerable attention, as a spectral-efficient multiple access technique [15]. Apart from spectral efficiency, NOMA is capable of increasing the number of served devices and also providing fairness among users. Thus, NOMA-based schemes are considered a promising alternative for the next-generation multiple access schemes, which are necessary in order to meet the connectivity requirements of 6G [16]. Due to the aforementioned capabilities, NOMA has the potential to reduce the communication cost during the WFL task. In this direction, a NOMA paradigm is investigated in the second of the two parts, where the *Compute-then-Transmit NOMA (CT-NOMA)* protocol is introduced and optimized. According to CT-NOMA the users terminate concurrently the local model training and then simultaneously transmit the trained parameters to the central server [17]. It should be highlighted that hybrid NOMA/OMA configurations are also worth of investigation, aiming to further reduce the latency and meet the WFL demands, by capitalizing on both the orthogonal and non-orthogonal aspects of multiple access and their underlying advantages. An example of the later considerations, could be the scheduling of few devices in the same orthogonal resource block, combined with the utilization of NOMA.

B. WFL over fog radio access networks

In order to satisfy the versatile requirements of 6G wireless networks, fog radio access networks (FRANs) have been proposed as a promising network architecture to provide low latency, massive connectivity and scalability. The edge nodes provided by the FRANs, which are empowered with powerful computation capabilities, could be an effective tool for assisting the training process during WFL. Firstly, FRANs can decongest the local devices from computationally-intensive tasks. Secondly, the provision of massive connectivity can lead to reduced communication cost, while it can also scale up the number of participating devices. Therefore, the FRAN is visualized as an intermediate layer between the participants and the cloud server, aiming to improve the training efficiency, by contributing with computational power and dealing with device density and connectivity issues.

A promising technique which leverages the aforementioned architecture is the Hierarchical Federated Learning (HFL) [18]. According to HFL, users update their local model and send the them to the fog nodes wirelessly. At this stage, the local models are aggregated by the fog nodes, i.e., a local averaging of the models is performed by the edge nodes. Following that, each fog node forwards the local averaged model to the central cloud via fronthaul links. Finally, the cloud server acts as a global aggregator and generates the global average model, which is reported to the users through the fog nodes, while the process is repeated until convergence. The considered multi-level configuration, could offer an efficient model exchange compared to the classical client-cloud architecture. However, the ubiquitous challenges of WFL, such as the straggler effect and device scheduling, still need to be resolved.

C. Asynchronous communication

Communication bottleneck is a significant burden for the WFL implementation, leading to increased delay, while 6G networks require ultra-low latency. In the conventional synchronous communication protocol, the latency of each round is determined by the slowest device. Thus, the latency per round and subsequently the model convergence speed is susceptible to the straggler effect. Furthermore, in case of clients' task completion failure, the progress of the model is wasted. To mitigate those phenomena, the asynchronous FL has been proposed as a promising solution [11]. The asynchronous configuration allows participants to join the FL task in an arbitrary time instant, even if a training round is still in progress. Hence, the server immediately updates the global modal after receiving a new local update by any user, justifying the term *asynchronous*, with respect to the model updates. The considered salient characteristic of asynchronous communication is representative of a practical WFL implementation, while it could further enhance the scalability of WFL. However, the bounded-delay assumption which is usually made, is unrealistic for practical FL systems. Therefore, novel asynchronous communication schemes should be developed, which also take into account users' behavior, the unreliability of wireless links and the efficient utilization of the radio resources.

D. Towards intelligent WFL implementation

In order to ensure an efficient FL implementation over 6G wireless networks, a plethora of performance metrics should

be considered and optimized. Hence, the network management and orchestration in a WFL setting, constitute an overriding, yet demanding issue to tackle with. More specifically, the optimization problems corresponding to the network control in WFL systems, such as resource allocation, power control, user scheduling, etc., are usually non-convex due to the coupling of several optimization variables. Moreover, the considered problems are often of a combinatorial nature. Thus, even by applying convex transformations, the solution might come along with high computational complexity, which is a major limitation and can significantly increase the overhead. Thus, conventional optimization techniques may be impractical for the realization of WFL. To alleviate this burden, machine learning techniques could be employed, to deal with non-convexity and retain low levels of computational complexity.

Deep Reinforcement Learning (DRL) is a auspicious machine learning technique, which could be exploited for the FL network optimization. It deals with agents who learn from the interaction with a dynamic environment. The agent aims to maximize a cumulative reward, which can refer to any figure of merit. Thus, the agent learns how to make better decisions, by observing the evolution of the considered reward. Therefore, DRL could be a powerful tool for resource management and decision making in WFL systems, aiming to solve complex problems in a near-optimal fashion and provide computationally tractable solutions.

Apart from leveraging ML techniques for enhancing the performance of WFL, federating learning itself, could also contribute as an intelligent tool for the self-optimization and orchestration of 6G networks. Therefore, WFL could leverage the data heterogeneity to model the network's behavior and optimize its performance through data-driven approaches, specially when privacy issues arise.

E. WFL in the next-generation internet-of-things (NGIoT)

WFL has the potential to directly contribute to most of the foundational challenges of the NGIoT, including reliability, energy sustainability, scalability, future-proof security and trust, privacy-by-design, etc [19]. To this direction, several interesting trade-offs can be investigated, such as between model accuracy and energy consumption at the mobile devices in the training process. On the other hand, one of the key challenges for the NGIoT is "the development of IoT data sharing and monetization enabling models and technologies" [19], which also remains a challenge when FL is used. Thus, the motivation and potential economic benefits from the participation in the training of the FL model, as well as the sharing of the model's output between different stakeholders is a particularly interesting topic. In this context, a promising ledger technology is the blockchain, which refers to a public and trusted ledger, operating on a peer-to-peer network without any third party being involved. Thus, WFL could benefit from blockchain in terms of security and privacy improvement, since it eliminates the need of a central server, while the participants are collaboratively building the global learning model through a consensus mechanism [20]. Finally, WFL can facilitate the construction of digital twin models of IoT devices in a distributed manner, which has also been recognized as a research priority for the NGIoT [19], [21].

REFERENCES

[1] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y. A. Zhang, "The roadmap to 6G: AI empowered wireless networks," *IEEE Commun. Mag.*, vol. 57, no. 8, pp. 84–90, 2019.

[2] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," *arXiv preprint arXiv:1610.05492*, 2016.

[3] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 1273–1282.

[4] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020.

[5] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings *et al.*, "Advances and open problems in federated learning," *arXiv preprint arXiv:1912.04977*, 2019.

[6] P. D. Diamantoulakis, V. M. Kapinas, and G. K. Karagiannidis, "Big data analytics for dynamic energy management in smart grids," *Big Data Research*, vol. 2, no. 3, pp. 94–101, 2015.

[7] Y. Wang, I. L. Bennani, X. Liu, M. Sun, and Y. Zhou, "Electricity consumer characteristics identification: A federated learning approach," *IEEE Trans. Smart Grid*, pp. 1–1, 2021.

[8] Z. Du, C. Wu, T. Yoshinaga, K.-L. A. Yau, Y. Ji, and J. Li, "Federated learning for vehicular internet of things: Recent advances and open issues," *IEEE Open J. Comput. Soc.*, vol. 1, pp. 45–61, 2020.

[9] D. Chen, L. J. Xie, B. Kim, L. Wang, C. S. Hong, L.-C. Wang, and Z. Han, "Federated learning based mobile edge computing for augmented reality applications," in *Proc. International Conference on Computing, Networking and Communications (ICNC)*, 2020, pp. 767–773.

[10] Z. Yang, M. Chen, W. Saad, C. S. Hong, and M. Shikh-Bahaei, "Energy efficient federated learning over wireless communication networks," *IEEE Trans. on Wireless Commun.*, vol. 20, no. 3, pp. 1935–1949, 2020.

[11] C. Xie, S. Koyejo, and I. Gupta, "Asynchronous federated optimization," *arXiv preprint arXiv:1903.03934*, 2019.

[12] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, and S. Cui, "A joint learning and communications framework for federated learning over wireless networks," *IEEE Trans. on Wireless Commun.*, vol. 20, no. 1, pp. 269–283, 2020.

[13] W. Shi, S. Zhou, Z. Niu, M. Jiang, and L. Geng, "Joint device scheduling and resource allocation for latency constrained wireless federated learning," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 453–467, 2020.

[14] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proc. ACM SIGSAC conference on computer and communications security*, 2016, pp. 308–318.

[15] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. K. Bhargava, "A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends," *IEEE J. Sel.Areas in Commun.*, vol. 35, no. 10, pp. 2181–2195, 2017.

[16] Y. Liu, S. Zhang, X. Mu, Z. Ding, R. Schober, N. Al-Dhahir, E. Hossain, and X. Shen, "Evolution of noma toward next generation multiple access (NGMA)," *arXiv preprint arXiv:2108.04561*, 2021.

[17] P. S. Bouzinis, P. D. Diamantoulakis, and G. K. Karagiannidis, "Wireless federated learning (WFL) for 6G Networks - Part II: The compute-then-transmit NOMA paradigm," *submitted to IEEE Commun. Lett.*, 2021.

[18] M. S. H. Abad, E. Ozfatura, D. Gunduz, and O. Ercetin, "Hierarchical federated learning across heterogeneous cellular networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 8866–8870.

[19] A. Brékine *et al.*, "Building a Roadmap for the Next Generation Internet of Things. Research, Innovation and Implementation 2021 – 2027 (Scoping Paper)," M. Brynskov, F. M. Facca, and G. Hrasko, Eds., Sep. 2019.

[20] D. C. Nguyen, M. Ding, Q.-V. Pham, P. N. Pathirana, L. B. Le, A. Seneviratne, J. Li, D. Niyato, and H. V. Poor, "Federated learning meets blockchain in edge computing: Opportunities and challenges," *IEEE Internet Things J.*, vol. 8, no. 16, pp. 12 806–12 825, 2021.

[21] Y. Lu, X. Huang, K. Zhang, S. Maharjan, and Y. Zhang, "Communication-efficient federated learning for digital twin edge networks in industrial IoT," *IEEE Trans. Ind. Informat.*, pp. 1–1, 2020.