Al-Quraan, M., Mohjazi, L., Bariah, L., Centeno, A., Zoha, A., Arshad, K., Assaleh, K., Muhaidat, S., Debbah, M. and Imran, M. A. (2023) Edge-native intelligence for 6G communications driven by federated learning: a survey of trends and challenges. IEEE Transactions on Emerging Topics in Computational Intelligence, (doi: 10.1109/TETCI.2023.3251404).

There may be differences between this version and the published version. You are advised to consult the publisher's version if you wish to cite from it.

https://eprints.gla.ac.uk/290444/

Deposited on: 30 January 2023

Enlighten – Research publications by members of the University of Glasgow
https://eprints.gla.ac.uk

# Edge-Native Intelligence for 6G Communications Driven by Federated Learning: A Survey of Trends and Challenges

Mohammad Al-Quraan, *Graduate Student Member, IEEE,* Lina Mohjazi, *Senior Member, IEEE,* Lina Bariah, *Senior Member, IEEE,* Anthony Centeno, Ahmed Zoha, *Member, IEEE,* Kamran Arshad, Khaled Assaleh, Sami Muhaidat, *Senior Member, IEEE,* Mérouane Debbah, *Fellow, IEEE,* and Muhammad Ali Imran, *Fellow, IEEE*

***Abstract*—New technological advancements in wireless networks have enlarged the number of connected devices. The unprecedented surge of data volume in wireless systems empowered by artificial intelligence (AI) opens up new horizons for providing ubiquitous data-driven intelligent services. Traditional cloud-centric machine learning (ML)-based services are implemented by centrally collecting datasets and training models. However, this conventional training technique encompasses two challenges: (i) high communication and energy cost and (ii) threatened data privacy. In this article, we introduce a comprehensive survey of the fundamentals and enabling technologies of federated learning (FL), a newly emerging technique coined to bring ML to the edge of wireless networks. Moreover, an extensive study is presented detailing various applications of FL in wireless networks and highlighting their challenges and limitations. The efficacy of FL is further explored with emerging prospective beyond fifth-generation (B5G) and sixth-generation (6G) communication systems. This survey aims to provide an overview of the state-of-the-art FL applications in key wireless technologies that will serve as a foundation to establish a firm understanding of the topic. Lastly, we offer a road forward for future research directions.**

***Index Terms*—5G, 6G, artificial intelligence, federated learning, wireless networks.**

## I. INTRODUCTION

Recent years have witnessed an unprecedented increase in the number of connected objects, which is attributed to the emergence of novel technological trends as well as the evolution of connected intelligence paradigms, promoting massive scale connectivity [1]. In specific, the number of internet-of-things (IoT) devices per human was 1.84 in 2010 with a total of 12.5 billion devices, while in 2020, this number increased to 6.58 devices per human with nearly a total of 50 billion devices [2]. With the remarkable revolutionary advancements in the field of wireless communications, it is envisioned that these numbers will continue to rise exponentially. Accordingly, enlarging connected devices, such as IoT, smartphones, industry machines, etc., will create a bottleneck on the limited resources of wireless networks. Therefore, there will be a continuous need to develop the existing network infrastructure to meet diversified demands.

According to the International Telecommunications Union (ITU) and the 3rd Generation Partnership Project (3GPP), the fifth generation (5G) wireless networks are designed to deliver improved quality of experience (QoE) by offering enhanced data rate, reliability, capacity, and energy efficiency. In light of this, 5G wireless systems were mapped out based on three fundamental concepts, namely, enhanced mobile broadband (eMBB), ultra-reliable and low latency communications (URLLC), and massive machine-type communications (mMTC) [3]. Nevertheless, the rise of services like extended reality and massive IoT, and the expected future applications such as holographic communications and multi-sense experience, impose far more stringent requirements than 5G networks and shed light on the next network improvements. Hence, the research will be shifted towards sixth-generation (6G) communication networks.

The fast-growing number of connected devices, coupled with the development of wireless communication infrastructures and the capability of embracing a wide range of intelligent applications, have resulted in unprecedented volumes of produced data traffic that need to be stored and processed; yielding the new concept of big data [4]. To harness the benefits of this data, artificial intelligence (AI), especially machine learning (ML), has become the cutting-edge technology that has the potential to exploit big data to deliver pervasive smart services and applications [5]. ML models are trained to perform diversified tasks by exploring hidden data patterns and drawing the value of such data to predict useful outcomes for several use cases, such as medical diagnosis and natural language processing [6].

In conventional ML algorithms, model training is performed centrally in cloud-based servers [7]. Datasets are collected and stored in one location, processed, and then employed to train ML models using one or multiple servers. This centralised

M. Al-Quraan, L. Mohjazi, A. Centeno, and A. Zoha are with the James Watt School of Engineering, University of Glasgow, Glasgow, G12 8QQ, UK, (e-mail: m.alquraan.1@research.gla.ac.uk, {Lina.Mohjazi, Anthony.Centeno, Ahmed.Zoha}@glasgow.ac.uk).

L. Bariah is with the Technology Innovation Institute, 9639 Masdar City, Abu Dhabi, UAE, (e-mail: lina.bariah@ieee.org).

K. Arshad and K. Assaleh are with the Artificial Intelligence Research Center (AIRC), College of Engineering and Information Technology, Ajman University, Ajman, UAE, (e-mail: {k.arshad, k.assaleh}@ajman.ac.ae).

S. Muhaidat is with the Center for Cyber-Physical Systems, Department of Electrical Engineering and Computer Science, Khalifa University, Abu Dhabi 127788, UAE, and also with the Department of Systems and Computer Engineering, Carleton University, Ottawa, ON K1S 5B6, Canada, (e-mail: muhaidat@ieee.org).

M. Debbah is with the Technology Innovation Institute, 9639 Masdar City, Abu Dhabi, UAE, (email: merouane.debbah@tii.ae) and also with CentraleSupelec, University Paris-Saclay, 91192 Gif-sur-Yvette, France.

M. A. Imran is with the James Watt School of Engineering, University of Glasgow, Glasgow, G12 8QQ, UK, and also with Artificial Intelligence Research Center (AIRC), College of Engineering and Information Technology, Ajman University, Ajman, UAE, (e-mail: Muhammad.Imran@glasgow.ac.uk)

nature of ML models limits their applicability for several emerging wireless network applications. The limitations include the following:

- Increased communication overhead between the end devices and the cloud resulting in network congestion and high energy consumption.
- The privacy is not by design, so security and data privacy are always a concern for conventional approaches.
- The propagation delay experienced in such ML techniques limits the implementation of centralised learning in real-time applications.

In light of this, federated learning (FL) has emerged as a promising solution to tackle the aforementioned challenges of centralised ML [8]. FL is a collaborative ML algorithm that uses distributed entities' datasets for local model training without the need to exchange any raw data with a central server. In FL, the role of cloud-based servers is limited to aggregating local models to develop a global model to be shared with all nodes in the network. Initially, a centralised server broadcasts initial model parameters to participating nodes, which leverage these parameters and their onboard resource capabilities for local model training. Next, once the training round is finished, each participant will send the local model updates to the FL server to aggregate the received local models. For enhanced accuracy, model training in FL is performed over multiple iterations; hence, after each training round, the server shares updated model parameters with participating devices to be utilised for the next training round. By using FL, the amount of data that needs to be sent to the server is reduced significantly, allowing only model updates to be sent to the server and hence, alleviating the pressure on the network resources. Furthermore, FL protects the endpoints' data privacy and security by allowing model training locally, where the data is generated.

### A. Related Surveys in the Literature

FL has attracted numerous interests and has been implemented in diverse applications across many areas. Notably, several surveys have been published since the advent of the FL algorithm. Table I summarises these surveys and highlights their significance. Here we outline the surveys in chronological order based on the publication date. The work in [9] categorises the FL systems into three categories, namely, vertical, horizontal, and federated transfer learning (FTL), and discusses the privacy techniques used in FL. The authors in [10] highlight the need to implement ML at the wireless network edge to facilitate reliable and low-latency communication. They explore the key building blocks of ML that allow for the transition from centralised, cloud-based model training to decentralised training techniques such as FL. Furthermore, a thorough investigation of the technical and theoretical frameworks of several case studies illustrates the importance of edge intelligence for beyond 5G (B5G) networks. Later, Kairouz *et al.* [11] introduce recent advances in FL by discussing techniques used to improve FL efficiency, explaining the methods used to preserve user data privacy, and how to make FL algorithms robust against attacks. The authors in [12] discuss possible threats and attacks, and highlight their implications to future FL algorithms, whereas [13] discusses FL properties and associated challenges in comparison to traditional distributed data centre computing. Du *et al.* [14] outline the importance and technical challenges of implementing FL in vehicular IoT networks.

Table I: Summary of Relevant FL Surveys.

| Ref. | Date | Authors | Article Main Topic |
|---|---|---|---|
| [9] | Jan. 2019 | Q. Yang, *et al.* | Categories of FL and Privacy Techniques |
| [10] | Oct. 2019 | J. Park, *et al.* | Edge ML in Beyond 5G Networks |
| [11] | Dec. 2019 | P. Kairouz, *et al.* | FL Advances, Problems and Challenges |
| [12] | Mar. 2020 | L. Lyu, H. Yu, Q. Yang | FL Threats and Attacks |
| [13] | May 2020 | T. Li, *et al.* | FL Implementation Challenges |
| [14] | May 2020 | Z. Du, *et al.* | FL Challenges in Vehicular Networks |
| [15] | July 2020 | M. Aledhari, *et al.* | FL Protocols and Enabling Technologies |
| [16] | July 2020 | V. Kulkarni, M. Kulkarni, A. Pant | FL Personalisation Techniques |
| [17] | July 2020 | W. Yang, *et al.* | FL in Mobile Edge Networks |
| [18] | Dec. 2020 | M. Chen, *et al.* | Collaborative FL |
| [19] | Jan. 2021 | Q. Li, *et al.* | Thorough FL Categorisation |
| [20] | Feb. 2021 | O. A. Wahab, *et al.* | FL in Communication and Networking Systems |
| [21] | Apr. 2021 | S. Abdulrahman, *et al.* | FL Architecture Extensive Explanation |
| [22] | June 2021 | L. Khan, *et al.* | FL Integration with IoT Networks |
| [23] | Dec. 2021 | Z. Yang, *et al.* | FL Implementation in Wireless Communications |
| [24] | Mar. 2022 | A. Z. Tan, *et al.* | Taxonomy of personalised FL |
| [25] | June 2022 | B. Ghimire, D. B. Rawat | Cybersecurity and FL in IoT |

The contribution in [15] spots the light on the concept of FL and illustrates some of the enabling technologies and recent research that addresses different FL perspectives. The study in [16] discusses the implications of training the FL model using heterogeneous datasets, and presents recent research that applies personalisation to overcome the data heterogeneity problem. While [17] is restricted in presenting the challenges associated with deploying FL in mobile edge networks only and provides the developed solutions that optimise these networks. Reliance on a central controller to organise the FL training process in IoT networks can limit the FL applications, and this issue is the authors' main focus in [18]. Accordingly, they have proposed a collaborative FL (CFL) framework where clients can implement FL with less dependence on the central server. CFL enables clients to engage in FL directly or indirectly, where some users are directly connected to the server while others are associated with neighbouring clients. Furthermore, this survey presents the original FL's architecture, benefits, and shortcomings compared to CFL. In [19], the authors present a thorough categorisation of FL in different aspects and discuss the existing solutions with their limitations in enabling FL. Abdel Wahab *et al.* [20]

present a tutorial on FL technologies and the associated challenges in communication and networking systems. The survey in [21] explains the FL architecture, system model and design, application areas, privacy and security, and resource management. The work in [22] presents a new taxonomy of FL in the context of IoT networks and explores FL's recent developments toward enabling intelligent IoT applications. Moreover, this survey introduces a set of metrics that can be considered when evaluating the performance of new FL algorithms. The review paper [23] highlights the requirements for FL in wireless communications, particularly for envisioned 6G systems. Besides, the motivation for using FL and the main obstacles accompanying FL implementation are discussed. Tan et al. [24] introduce the key motivation and the taxonomy of personalised FL, which is the technique used to handle the statistical heterogeneity of real-world datasets to learn ML models collaboratively. The authors in [25] study the use of FL in cybersecurity and vice versa, and discuss several approaches that address IoT networks' performance issues when deploying FL.

### B. Contributions

It is worth emphasising that to the best of our knowledge, no prior works presented a comprehensive study of FL potentials and applications for various existing/next-generation wireless networks. Besides, most of the aforementioned survey papers generally focus on specific technological trends or aspects associated with FL applications. Conversely, in this survey, we provide a systematic review with a featured presentation that leads the reader to a thorough understanding of the FL algorithm and its recent advents, as well as its envisioned implementations in various types of B5G/6G wireless networks. Moreover, this article offers numerous future research opportunities derived from the latest trending technologies that have not been covered by any previous surveys to the best of the authors' knowledge. The following points demonstrate our main contributions:

- We present a concrete conceptual background on the working principles of the FL algorithm. Also, we describe its architecture, categories, operation, and optimisation schemes.
- We explore the enabling technologies that create the stepping stones for facilitating the operation of FL.
- We provide an in-depth discussion of the key drivers for deploying FL in state-of-the-art wireless applications, taking into account the associated performance metrics and ongoing research. Moreover, we discuss the vision for integrating FL with new potential prospective areas in future wireless networks.
- The survey delves into highlighting the challenges associated with the operation of FL in emerging wireless technologies, and identifies the approaches proposed to tackle them.
- We offer a look ahead towards unexplored possibilities drawn from modern technology trends to reap the benefits of FL implementations in the context of cutting-edge future research directions.

It is noteworthy that the survey structure is organised and written in a distinct taxonomy to make it easier for the reader to navigate and recognise the contributions made in each area.

### C. Organisation

The rest of this paper is organised as follows. Section II introduces the fundamental aspects of the FL framework covering architecture, categories, operation, and aggregation schemes. Then, we give the key enabling technologies of FL in Section III. Section IV presents a comprehensive study of FL applications in various wireless networks. In addition, this section discusses its applicability in new potential areas of B5G/6G networks. After that, the FL challenges and corresponding mitigating techniques are outlined in Section V. Section VI points out future research directions from different perspectives. Finally, Section VII gives concluding remarks. Fig. 1 illustrates the detailed outline of this survey.

## II. PRELIMINARY: FL FUNDAMENTALS

The concept of FL has attracted significant attention in academia and industry [26]. The key principle of FL is to construct a generalised global model by performing distributed model training. The recent advancement in edge devices' communication and computation capabilities and the large amount of data generated and stored locally on the devices facilitate the spread of this emerging technology widely. This section presents the fundamentals, architecture, categories, operation principles, and aggregation schemes of FL algorithms.

### A. FL Architecture

Based on the nature of the network, the architecture of FL can be categorised into classical and hierarchical FL (HFL). The classical FL approach consists of two main parts: the server and the participating clients [8], as illustrated in Fig. 2(a). The FL server must have certain specifications to orchestrate the FL process efficiently. These specifications are drawn from the considered ML technique and the number of clients. For instance, training a deep learning (DL) model through many clients requires a high server capacity, a huge computation capability, high-speed interfaces, and locating the server in close proximity to the clients. On the other hand, the specifications may be less stringent when considering simpler models of neural networks and a few clients. At the beginning of the FL process, the server will initiate the training procedure by sharing a new or pretrained model with the participating clients. After that, the clients will personalise the received model by training it based on their local data, and then share their local models with the server for aggregation and global model update. On the other hand, HFL framework [27], depicted in Fig. 2(b), optimally fits in heterogeneous networks that include different cell coverage. This architecture is introduced to alleviate the bandwidth (BW) overhead at the FL servers, resulting from the large number of model updates communicated from the clients. Furthermore, HFL can reduce the communication latency experienced between the clients and the server by reducing the link distance. The HFL framework consists of two stages; in the first one, the clients

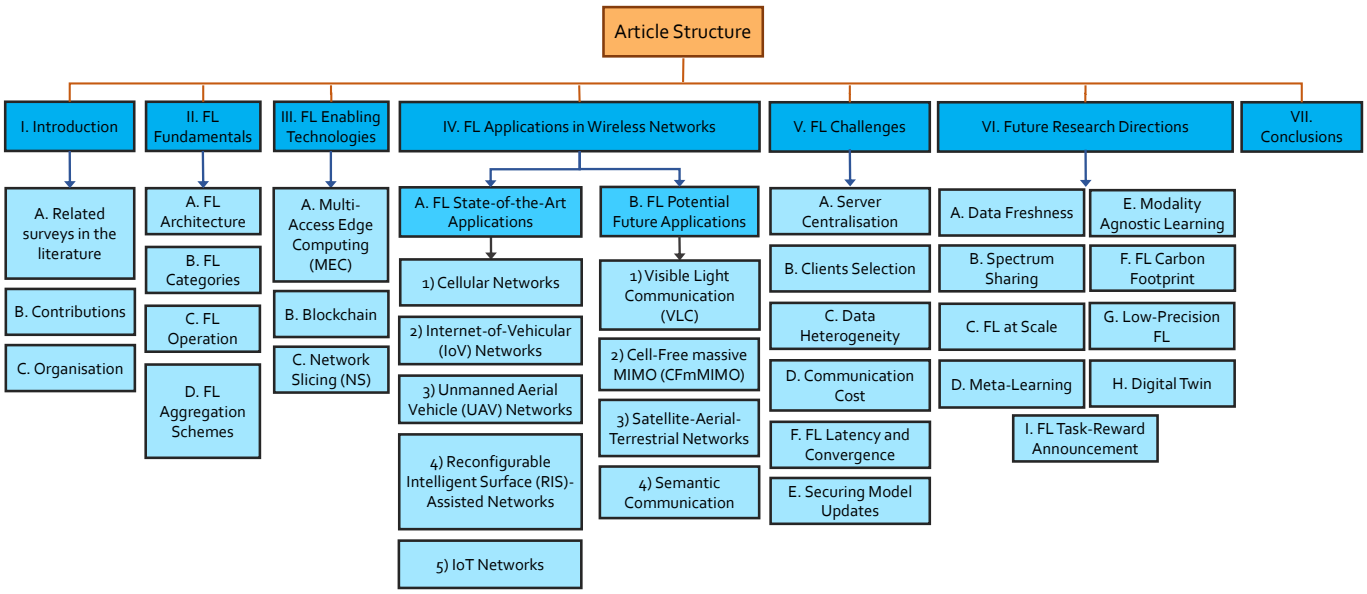


Figure 1: Illustrative diagram of the paper structure.

send and receive the model parameters by communicating with a server located at the small base station (SBS), i.e., the edge server, and the server performs local model aggregation. Meanwhile, in the second stage, the edge servers send the aggregated models to a central server that can be located at the macro base station (MBS) or in the cloud, in which the server performs edge model aggregation for global model update and sends it back to the edge servers.

It is worth mentioning that the need for robust communications between the clients and the FL server, mainly when the latter is located in the cloud, is mandatory to guarantee a seamless FL training process. However, the current internet links' capacity is insufficient to meet the emergency demands, along with the growing connectivity needs from different sectors, such as industry, education, and transportation. This results in the need to move to the new concept of worldwide decentralised internet, which can be achieved using decentralised mesh networks [28]. Such networks rely on establishing connections between different nodes, i.e., consumers and businesses, to make alternative ways of connectivity other than the known centralised internet service provider connection. Decentralised mesh networks are reliable for maintaining the connections between the participating clients and the FL server and ensuring a smooth training process.

## B. FL Categories

Given the significant role of local datasets in realising efficient training and assuming the data is maintained in a 2D matrix form, rows represent data samples, and columns indicate features. FL systems can be classified based on the data distribution characteristics between different parties into horizontal, vertical, and FTL [19].

*1)* ***Horizontal FL****:* This is the most common category of FL, also called sample-based FL. The unique characteristic of this category is that the datasets of different parties share

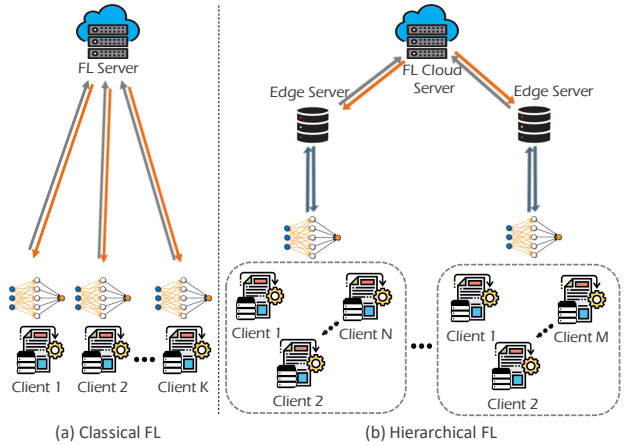


Figure 2: Types of FL architecture (a) Classical FL in client-server architecture (b) Hierarchical FL in client-edge-server architecture.

the same feature space while differing in the sample space. For example, two regional educational institutes have similar interests in monitoring the research outcomes, representing the feature space, while they have different research groups that denote the sample space. This category facilitates the adoption of a unified ML model with the same architecture for all datasets. Therefore, the global model can be obtained by averaging all local updates. FedAvg technique [29] is an example of this type of FL system.

*2)* ***Vertical FL****:* This category, referred to as feature-based FL, can be exploited when two or more datasets share the same sample space while their feature spaces are distinct. For instance, considering two different parties in the same city, one is a healthcare institute, whereas the other is an e-commerce company that records the customers' buying habits. Their user sets are most likely to have residents from that area, which means the same sample space. The objective here is to exploit the different features of these two parties to build a

model that predicts the future health status of the residents based on their buying practices. When implementing vertical FL, the participating parties may be curious to know each other's data, so a trusted third-party coordinator can protect the data confidentiality during the training process. However, if a certain level of trust exists between the participating parties, the need for a third party can be eliminated, and one of the parties can be the coordinator.

*3) **FTL**:* When the dataset of different clients slightly intersects in the feature and sample spaces, FTL (or hybrid learning) is the best candidate. FTL enables knowledge transfer from one domain to another, which helps in achieving better learning results. Specifically, a local model trained in one party is transferred to another party to leverage information extracted from the non-overlapping regions for enhanced model training at the other party. The most common example of transfer learning is the image classification problem. Several models exist that are tailored for classifying specific datasets, and they can be used to classify other types of datasets after making a minor tuning.

### C. FL Operation

FL protocol consists of three main phases [30] detailed as the following:

1. **Clients selection**: Albeit large-scale deployment is an attractive feature in FL, compared to classical ML, the number of clients participating in model training can easily reach thousands or millions of devices. This enormous number of endpoints reflects the capacity enhancements anticipated to be delivered by B5G (1 million/km$^2$) and 6G networks (100/m$^3$). As a result, end-device onboard capabilities and data distribution will vary considerably among the participants, rendering client selection a critical design aspect in FL. Several methods are proposed to address this issue, such as [31], where the authors propose a technique that improves the time-to-accuracy training performance by guiding the FL developers to select participants even at the scale of millions of clients. Further approaches are discussed in Section V-B.

2. **Configuration**: In this phase, the selected participants receive the initial model parameters and train their local models based on the local datasets. In particular, after selecting participating devices successfully, $K$ edge nodes are ready to begin the training process. Fig. 3 illustrates the FL's architecture and the operation steps. The device, $k \in \{1, 2, ..., K\}$, has a local dataset, $D_k \in \{D_1, D_2, ..., D_K\}$, which includes input-output pairs of samples $(x_i, y_i)$, $x_i, y_i \in \mathbb{R}$. In step ①, the FL server initiates the global model created to perform a specific task and shares it with the selected participants. Next, at the $t$-th iteration, each participating node acquires the model weights $W_{t-1}$ and begins the model training by exploiting the data samples on their local storage. The objective of model training is to minimise the loss function $F_k(W_t^k)$ of all data samples in the training dataset, $F_k(W_t^k) = \frac{1}{D_k} \sum_{i \in D_k} f_i(W_t^k)$, i.e., obtaining the optimum model parameters $W_t^k$ that minimise the loss function at each round of training which can be represented mathematically as, $\arg\min_{W_t^k \in \mathbb{R}} F_k(W_t^k)$. Where

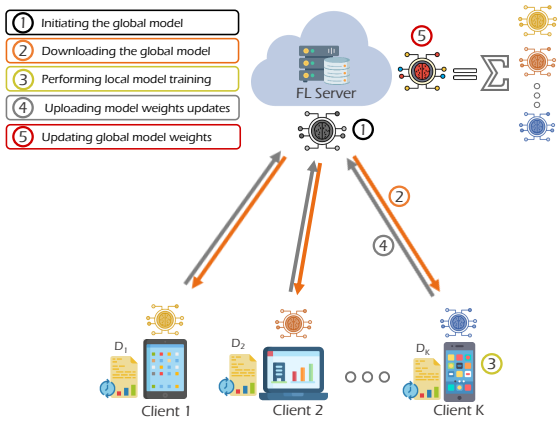


Figure 3: Sequential operation steps of FL involving $K$ participants.

$f_i(W_t^k)$ indicates the loss on data sample $i$ given the parametrisation $W_t^k$, (steps ②, ③).

3. **Reporting**: At this point, the participants share the local model updates with the central server in a synchronous, or asynchronous manner [32]. Finally, the server aggregates these models to update the global model. Specifically, models aggregation and global model parameters computation are performed at the server as the following $W_t = \sum_{k=1}^{K} \frac{D_k}{D} W_t^k$, where D represents the entire dataset of all clients, i.e., $D = \sum_{k=1}^{K} D_k$, (steps ④, ⑤). The steps from ② to ⑤ are repeated until the global model converges to a desired accuracy.

### D. FL Aggregation Schemes

Gradient descent (GD) algorithm that aims to find the minimum of a differentiable function is commonly used in various ML algorithms, especially NN models [33]. However, the computational complexity of GD increases with the dataset size, making it unsuitable for FL systems due to the slow convergence rate. An alternative to GD is stochastic GD (SGD), which can perform gradient calculation over a subset of data, significantly enhancing the convergence rate. In the FL setting, SGD (FedSGD) is exploited as an approach to quantify how often the global FL model needs to be updated [29]. FedSGD is the basic aggregation scheme for FL-enabled systems, where clients compute gradients using random data samples. However, the FedSGD technique requires many communication rounds proportional to the volume of nodes' datasets, which will burden the communication links and consume BW.

To address the above problem, the federated averaging (FedAvg) strategy has been proposed to alleviate the pressure on communication resources [29]. FedAvg is a generalisation of FedSGD, where each node repeatedly runs SGD locally over different local data subsets and finds the optimum model parameters by averaging the locally evaluated gradients. Three main parameters control the performance of FedAvg: (i) the fraction of the selected nodes that perform computation at each round, (ii) the size of data subsets, and (iii) the number of epochs that the node passes over its dataset in every round. In FedAvg, instead of sending the computed gradients, each

node will only send the model parameters. Thus, compared to FedSGD, the FedAvg algorithm performs more local computation and less communication with the server.

Nevertheless, in real-world scenarios, in which network devices are heterogeneous and the local datasets are non-identically distributed, FedAvg experiences poor convergence behaviour. Therefore, some variants of the FedAvg algorithm have been introduced to develop faster aggregation techniques. FedProx was proposed to solve the heterogeneity issue in federated networks [34]. The FedProx principle is similar to that of FedAvg but with a small critical modification that improves performance. Instead of forcing every node to perform the same computation work, FedProx considers system heterogeneity by allowing each node to perform an amount of local computation proportional to its resources. Accordingly, enabling parameter aggregation from a set of heterogeneous nodes.

Another extension to the FedAvg scheme is the FedSplit algorithm [35], which relies on the operator splitting procedure for convex optimisation problems. Operator splitting is an efficient method for solving large-scale convex problems by performing iterations of simple and computationally inexpensive operations. It converts the problem into simpler sub-problems and makes progress on them separately. Motivated by the failure of FedAvg and FedProx to preserve the fixed points of the original optimisation problem, FedSplit is proposed as a splitting algorithm for federated optimisation to achieve rapid convergence. Moreover, the work in [36] applied the adaptive optimisers ADAGRAD, ADAM, and YOGI in the FL setting, i.e., FedAdaGrad, FedAdam, and FedYogi. Extensive experimental evaluations are performed to examine these algorithms compared to the FedAvg algorithm. Furthermore, the Qsparse-local-SGD algorithm [37] considers both local computation and communication reduction in distributed settings. Convergence analysis is made in synchronous and asynchronous FL, showing that the Qsparse-local-SGD algorithm achieves the same convergence rate as FedSGD.

The above approaches are mainly designed for NN models where the parameters, i.e. weights and biases, are the main elements to update the global model. Despite numerous attempts to enhance the aggregation process, NN and DL models incur high communication and computation costs. Therefore, several studies have begun to explore other low-complexity techniques, such as ensemble learning under the FL settings, like FedBoost [38] and FedTrees [39]. It has been demonstrated in [38] and [39] that when the federated model is trained according to these algorithms, an excellent performance is achieved in terms of accuracy, computation time, and communication rounds. This paves the way for exploring other ML techniques in the FL environment.

## III. FL Enabling Technologies

With the aim to realise the full potential of FL, several enabling technologies can be leveraged in order to improve the performance of FL and hence, accentuate its promising features. This section is devoted to discussing some of these enabling technologies.

### A. Multi-Access Edge Computing (MEC)

The rapid evolution of the internet-of-everything (IoE) paradigm has resulted in a plentitude of end devices. The abundance of resource-intensive devices, coupled with the emergence of QoE-oriented applications, has led to a wealth of data being generated at the edge of wireless networks. Exploiting this data requires sending it across the networks to reach the cloud server where the significant computation and storage resources are located. Accordingly, cloud computing has become unsuitable for resource-limited real-time applications, owing to the increased overhead occurring in the network, in terms of energy and spectrum resources, in addition to the increased latency and compromised security. Therefore, the European telecommunications standards institute (ETSI) has introduced a new computing paradigm called MEC, which brings cloud computing capabilities to the edge of the radio access network [40]. The key motivation behind MEC is that running applications closer to the end-users with their associated computation tasks will reduce network congestion and preserve the network resources, enabling enhanced user experience.

The proliferation of smart end devices with the employment of MEC provides a suitable environment for employing FL algorithms. Shifting to decentralised ML model training at the network edge allows for greater scalability by distributing the computation from centralised architectures of the network core/cloud to the edge closer to the users. Moreover, MEC enables FL algorithms to offer latency optimisation for real-time applications where data aggregation, analytics, and computation are handled within user proximity. In fact, the capabilities of the edge server enable it to act as an FL server, whereas the widely dispersed edge devices are used as FL clients. Thus, MEC and FL provide rich services and applications close to the end users.

### B. Blockchain

As a decentralised learning algorithm, FL has benefits in two main aspects: load balancing and privacy-preserving. However, FL has shortcomings as it does not keep records of participants' training contributions along with reliance on a central server prone to a single point of failure. In this regard, blockchain, a decentralised database managed by distributed nodes, can play an essential role in FL. Blockchain was initially introduced in 2009 as a type of distributed ledger technology [41]. In particular, blockchain was primarily proposed to serve as a ledger of the public transactions for the cryptocurrency Bitcoin. For improved security, the principle operation of blockchain relies on grouping multiple transactions and storing them in a block encrypted by a hash signature. After that, each new block is time-stamped and chained with the previous one, creating a long chain of encrypted chronologically-ordered transactions. Therefore, blockchain has a high level of security, as altering the content of any block requires an agreement from all nodes connected to the chain. These merits motivate the authors in [42] to design an incentive mechanism for a blockchain-enabled FL platform that can record and secure the workers' updates and reward them accordingly. Whereas

the study in [43] sheds light on the distributed ledger feature of blockchain to realise decentralised FL training without needing a central server, and proposes a new paradigm called FLchain.

Furthermore, the blockchain provides a fully transparent network in which all nodes can observe all transactions coming in and going out. When a new transaction is stored in the blockchain, it is considered immutable because it is verified based on a consensus mechanism. The consensus mechanism validates the data in each block and verifies its availability since all blocks will store the same copy of the data. This has been exploited in [44] when the authors designed a blockchain-based FL system that can prevent malicious model updates using blockchain's immutability and decentralised trust property. Moreover, the work in [45] uses blockchain to support the operation of the FL utilised to provide up-to-date service provisioning and support device communication in vehicular environments. Blockchain technology is adopted to ensure the credibility and integrity of sensitive services, such that the local models are verified using a consensus algorithm. Similarly, the authors in [46] integrated blockchain with FL to maintain and secure local model parameters, improve learning quality, and optimise the allocation of varying resources in B5G networks. In light of the above discussion, decentralisation, availability, transparency, immutability, and security are the most promising features of the blockchain, which are well-suited for the FL system and constitute one of its enabling factors.

### C. Network Slicing (NS)

NS is one of the key enablers in B5G wireless networks, where it exploits the network's physical structure to create several independent logical networks called slices [47]. Each slice comprises an end-to-end isolated network tailored to fulfil diverse application requirements. In this respect, millimetre wave (mmWave) and terahertz communications (THz) in B5G/6G networks enable improved capacity for devices operating in a small coverage area, allowing the realisation of different IoE networks. These networks will require resources that meet the diverse quality-of-service (QoS) requirements. The unique characteristic of NS is that it grants each network segment an isolated and tailored slice to enable a particular service. However, configuration, activation, association, and management of network slices constitute a challenging factor that requires developing dedicated intelligent techniques. Therefore, using AI is a must to optimise real-time resource allocation and distribution among different slices according to their requirements. In light of this, FL and NS are considered promising enablers for each other. The work in [48] presents an FL-based framework that predicts slices' service-oriented key performance indicators (KPIs). The concept of an in-slice manager was introduced for monitoring and collecting slices' KPIs and local decision-making to ensure optimal performance.

NS allows future mobile communications to ensure the efficient allocation of services while guaranteeing the QoS. For this reason, the study in [49] proposes an FL-based forecasting algorithm to predict base station level traffic in sliced network architecture to facilitate intelligent and predictive management of resources. Whereas the authors in [50] present a federated deep reinforcement learning (DRL) scheme to manage the transmission power and spreading factor resources in LoRa-based industrial IoT (IIoT) slices. A multi-agent self-model is trained under the FL environment to obtain an optimal decision of LoRa parameters that fulfil the QoS of IIoT virtual network slices. Furthermore, the proposed work in [51] offers a hybrid federated RL framework to find the optimal device association for radio access network (RAN) slices to maximise the network throughput.

## IV. FL APPLICATIONS IN WIRELESS NETWORKS

Since the advent of FL by Google in 2016, extensive research has been conducted to promote, enhance, and determine the best usage of this decentralised learning algorithm. Wireless networks are one of the forerunners to adopt FL in their architecture, as depicted in Fig. 4. This section will thoroughly present the key driving applications of FL in wireless networks; more specifically, Section IV-A sheds light on the existing FL applications and accompanying challenges along with their potential solutions. Whereas Section IV-B describes the significance of FL in new and promising application areas of the forthcoming B5G and 6G networks.

### A. FL State-of-the-Art Applications

This section presents the research that has been done on utilising FL in current wireless networks.

#### 1) Cellular Networks:

The rollout of 5G in late 2020 has allowed operators to launch numerous commercial services that benefit from the enhanced features provided by this new technology [52]. In addition, the use of FL in these networks has many applications in different areas, as described below.

***Homogeneous Cellular Networks:*** This type refers to low-frequency wireless networks with macrocells, alluding to their wide coverage. Two main concerns for FL at the network edge are heterogeneous devices with different computation and communication capabilities and securing local model updates. The work in [46] presents a blockchain-enabled FL framework to ensure security in a trustless environment using a distributed ledger between entities. Blockchain is an intermediary between the FL server and edge nodes to verify model parameters based on the consensus process. Also, FL has applications for network function virtualisation (NFV), which is introduced as an innovative concept that enables adaptive resource allocation for future wireless networks. Subramanya *et al.* [53] leverages the FL technique to build a model that can proactively predict the auto-scaling setting for MEC virtual services and ensure data protection policies.

***Heterogeneous Cellular Networks:*** Heterogeneous networks (HetNets), which comprise different cell types, expand wireless networks' coverage and capacity. FL can be implemented in HetNets for resource allocation purposes. It was demonstrated in [27] that applying HFL by grouping the users and assigning the needed resources for transmission can reduce the end-to-end communication latency in HetNets. This can

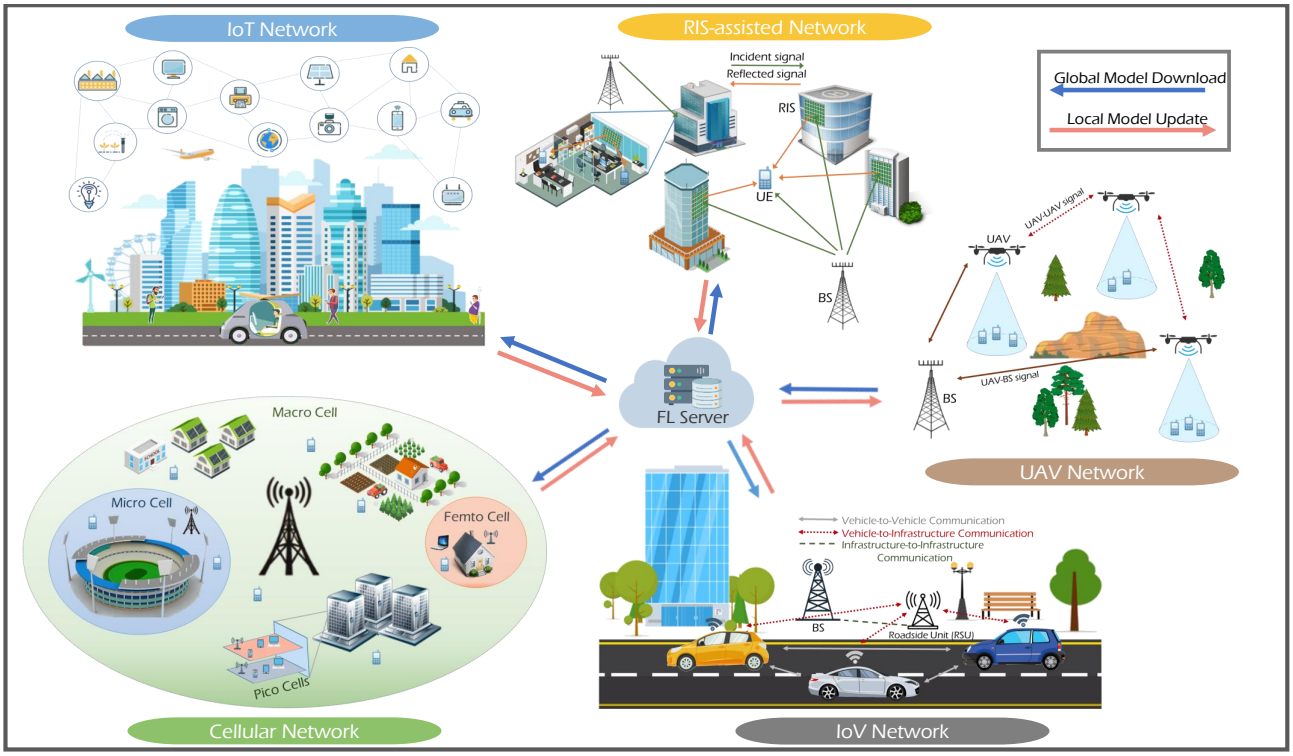


Figure 4: FL in various wireless networks; FL algorithm in the context of single or multiple wireless networks.

be achieved by dividing users into clusters and assigning each cluster to the closest SBS. On the other hand, 5G HetNets are vulnerable to attacks, like denial of service (DoS), evil twinning, and port scanning. The work in [54] proposes a node-edge-cloud framework empowered by HFL to detect attacks throughout the 5G HetNets. Specifically, multiple dedicated nodes are distributed inside the network, each of which performs model training by employing the RL technique to enable adaptive learning that can capture the rapidly changing nature of the HetNets environment. In addition, the work in [55] presents FL-empowered MEC framework to tackle the communication overhead and delay between the edge server and clients in FL to enhance the training efficiency.

***Multiple-Input Multiple-Output (MIMO)****:* FL has many applications associated with MIMO technology. Given the high dynamicity of mmWave systems, the study in [56] enhances the performance of massive MIMO systems by estimating channel state information. FL is leveraged to conduct decentralised learning on the user side using local pilot signals to predict channel matrix, which helps determine the best beamforming design and improve the system's performance. Moreover, the work in [57] presents an energy-efficient solution to support multiple FL groups in future wireless systems. Massive MIMO is utilised to assist model updates and ensure a stable operation of multiple FL processes executed within the same coherence time.

***Fog-Cloud RAN****:* The ever-increasing number of connected devices in 5G and beyond networks necessitates the transition to an ultra-efficient air interface. As a result, two air interface structures evolved, namely cloud-RAN (CRAN) and fog-RAN (FRAN). When surveying the literature, we observed many FL applications in FRAN networks, but using FL in CRAN networks is scarce. For example, the study in [58] optimises the latency and BW resources when deploying FL in reconfigurable intelligent surface (RIS)-aided CRAN systems. The RIS controls channel propagation conditions and supports over-the-air computation (AirComp) technique to perform coherent on-air aggregation for local models by allowing simultaneous transmissions from clients to the parameter server. On the other hand, the FRAN paradigm fully uses edge networks and provides vital features such as content caching for optimal application performance and user experience. The authors in [59] propose an FL-based mobility-aware content-caching framework in FRAN-based networks. Mobility and content demand statistics are exploited to improve users' QoE by predicting and caching the most likely future content.

***5G New Radio (5G-NR)****:* 5G-NR is a new radio interface standard designed by 3GPP to satisfy the growing demands of 5G mobile networks. This new radio access technology allows user equipment to switch dynamically between different resource blocks with different BWs. However, such a technique raises resource allocation challenges in B5G networks. FL has many applications in resource allocation in terms of computation, communication, and energy efficiency. For example, the study in [60] uses FL to develop an ML model that aids in performing distributed resource management in cellular networks while minimising uplinks transmit power.

*2) Internet-of-Vehicular (IoV) Networks:*

IoV has recently emerged as a key enabler for intelligent transportation systems (ITSs), combining two key concepts, namely, vehicle networking and intelligence [61]. Within this context, the IoV paradigm aims to achieve smart information

interaction between a vehicle and all network entities. Whereas vehicle computation capabilities realise vehicular intelligence by exploiting DL algorithms, cloud and edge computing, and big data analytics.

***FL in ITS**:* Communication reliability and latency are particularly significant in ITSs, owing to the severe consequences that might affect human safety. The proposed work in [62] exploits the integration of FL with blockchain to realise a distributed, privacy-aware, and efficient model designed for autonomous vehicular networks. The diverse nature of the vehicles in ITSs is particularly appealing for FL applications. The heterogeneous data helps improve the model accuracy by incorporating all network scenarios experienced by different vehicles. In addition to latency, FL has shortcomings in server centralisation, where exchanging large updates between the participants and the server yield a high overhead on the server. To overcome this challenge, [62] employs the blockchain technique, in which the distributed ledger is shared with each vehicle and maintains copies of the global and private models available and verified by each vehicle, relieving the pressure imposed on the central server. [63] studies the use of FL setting within the context of URLLC in vehicular networks. It mainly focuses on proposing a distributed joint transmit power and resource allocation framework that can reduce the power consumption of vehicular users while ensuring low-latency communications.

***Vehicular Edge Computing (VEC)**:* Following a similar concept to MEC, VEC exploits the communication and computation capabilities at the network edge. Ye *et al.* [64] implement FL with VEC to perform image classification to support diverse applications in ITSs. A model-selective approach was proposed to select clients with the highest computational capabilities and select models with the best image quality for aggregation. In an asymmetric FL setting, the server has no information about clients' data and resources. To this end, a two-dimensional contract mechanism is proposed in [64], in which the server designs contract bundles that include various levels of data quality, computation capability, and rewards, and then the clients select the bundles that increase their utility. As part of IoV networks, electric vehicle (EV) networks are becoming more popular as the number of EVs increases; such networks are expected to take over from traditional vehicles in the coming years. The work in [65] studies energy efficiency and profit maximisation at charging stations (CSs). It proposes an FL-based economically efficient framework to investigate the historical energy transactions to increase CSs profit. Specifically, FL is used to train a local model using CS private data to predict the EVs' energy demands. After that, the local models of every CS are aggregated and shared amongst them to benefit from other CS information, yielding more accurate results.

***Traffic Perdiction**:* Traffic prediction in smart cities brings up many benefits for ITSs, such as road safety, congestion avoidance, and shortest route selection. These gains are pronounced when exploiting information gathered from the edge in parallel with FL. One enhancement technique for FL is selecting the best hyperparameters of the local models in edge devices. As most of the literature focused on FL global optimisation, privacy, and communication, very few studied optimising model parameters. Qolomany *et al.* [66] proposed a particle swarm optimisation (PSO)-based technique to optimise local hyperparameters at the edge devices. Specifically, PSO optimises the local NN parameters, including the number of layers, neurons per layer, and epochs. This optimisation technique has been evaluated in traffic prediction as a use case. The work shows that the number of client-server communication rounds to find the best parameters is significantly reduced. This technique is attractive due to its low complexity implementation. However, its limitation lies in the reliance on a random search for the best initial parameters, which requires an unpredicted time that may affect the whole learning process.

*3) Unmanned Aerial Vehicle (UAV) Networks:*

The flying vehicles in a UAV network have many attractive features, such as low cost, mobility flexibility, and ease of deployment, enabling them to participate in many tasks considered hard to perform. The application of AI algorithms and the recent advancements in UAV technology have widened the use-cases ambit of UAV networks [67].

***AI-empowered UAV**:* The interplay between AI and UAV networks opens a new horizon for exploiting UAVs in more complicated tasks; however, data security and privacy remain significant challenges. In UAV-enabled mobile crowdsensing (MCS) applications, FL is particularly appealing for preserving the privacy of sensed data. In this regard, the authors in [68] integrated an FL-based UAV network with blockchain technology to eliminate the need for a central server. In addition, blockchain enhances FL network security by expulsing the adversary clients and sharing safe model updates between clients. On the other hand, the work in [69] proposed an FL-enabled air quality monitoring framework for secure MCS. A UAV swarm is utilised to measure the air quality, and the sensed data is used to train a lightweight model to predict the air quality index. FL is considered a promising candidate that can exploit the data silos collected by different agencies to produce a global model while preserving data privacy.

Following the MEC concept, federated edge learning (FEEL) can potentially reduce the end-to-end latency and communication overhead in UAV networks. Yet, as demonstrated in [70], the efficient implementation of FEEL in UAV-based IoT networks is restrained by the battery lifetime of UAVs. In this respect, computation resource and BW allocation optimisation were formulated in [70] to enhance the FEEL performance in a UAV network. Also, in [71], FL has been utilised as an aided technique to reduce the communication cost between multiple UAVs and a ground fusion centre in the context of image classification for remote area exploration missions.

***Flying Ad-hoc Networks (FANETs)**:* With the interest of accomplishing complicated tasks in UAV networks, UAVs are grouped in an Ad-hoc manner to create a local network, which allows UAVs to cooperate to perform joint tasks. Recent trajectory design and remote monitoring developments rely primarily on ML algorithms [72]. To recall, such classical

algorithms do not fit in the context of UAV networks due to their high mobility and constrained energy resources. FL was proposed to reduce the communication overhead as an efficient paradigm for FANETs, in which all participating UAVs collaborate to estimate the initial model parameters. Then, initial model parameters from all UAVs are shared and leveraged for local model training. A FEEL server is employed for model aggregation to exploit the local models to develop an enhanced global model.

Attributed to the inherent non-centrality nature of FANETs, such networks are vulnerable to several security threats that intend to disrupt their functionality, such as impersonation and jamming attacks [73]. Centralised attack detection and mitigation approaches are impractical, owing to the highly dynamic topology of FANETs. Thus, decentralised techniques are mandatory for such types of networks. To this end, in [74], an FL-based device jamming detection for UAVs in FANETs was proposed. In addition to the enhanced security, the framework in [74] has considered the data heterogeneity issue between different UAVs. In particular, a Dempster-Shafer technique categorises UAV clients based on their data quality into groups. Then the FEEL server selects high-quality data group(s) for model training purposes.

*4) Reconfigurable Intelligent Surface (RIS)-Assisted Networks:*

The emergence of numerous mmWave and THz applications has flagged several concerns attributed to the vulnerability of such applications to signal blockage and shadowing effects. Motivated by this and with the recent advancements in the solid-state industry, RISs have emerged as enablers of future wireless networks [75]. An RIS, comprising several reflective elements, can be artificially engineered to control the electromagnetic properties of wireless signals and enable diverse functionalities, including wave splitting, reflection, absorption, etc. Leveraging an RIS is particularly beneficial in AirComp-enabled FL scenarios, in which some clients may be experiencing blockage or weak channel conditions, affecting the global model training quality [76]–[78]. AirComp is a technique that exploits the superposition nature of the wireless channel to transmit simultaneous model updates from multiple clients. Section V-D covers the details of this technique. Yang *et al.* [76] use the AirComp technique assisted by RIS to boost fast global model aggregation, which reduces the required radio spectrum for parameter transmission since the clients collectively send their updates using the same channel. Also, to further enhance and boost the global model aggregation quality, an RIS is used to reduce aggregation errors by strengthening the quality of combined signals. In this respect, aiming to unleash the full potential of RIS in FL settings, Liu *et al.* [77] formulated a joint communication and learning optimisation problem by taking into consideration device selection, transceiver design, as well as RIS parameters.

The aforementioned contributions have assumed perfect channel state information (CSI) at the server and clients' sides. However, acquiring CSI at the transmitter (CSIT) is not always attainable due to dynamic channel conditions, leading to a significant delay in receiving the CSI information, thus curbing the FEEL global model convergence. The proposed work in [78] investigated the CSIT-free over-the-air model aggregation based on RIS-assisted FEEL. The CSI at the transmitter side is assumed to be unavailable, while perfect CSI is assumed at the server side. Besides, the RIS adjusts and aligns the channel coefficients with the model aggregation weights. To this end, the successive channel coefficients are constrained, as a function of RIS phase shifts, to be proportional to the weights of the local models. Moreover, the received scaling factor is optimised by minimising the aggregation mean square error. To solve this optimisation problem, a difference-of-convex algorithm was adopted. Furthermore, RIS has proven its efficiency in converting wireless channels into a smart electromagnetic environment. To realise high-speed RIS-based communication, the authors in [79] proposed two FL-based RIS optimisation schemes: RIS-assisted outdoor and indoor IoT mmWave communications. In the former scenario, the RIS controller is considered the FL server, while the user equipment (UE) is a client. The clients' data represents the CSI corresponding to their location and optimum RIS configuration. The trained model is aimed to optimise the achievable rate to enable high-speed mmWave communications. The latter scenario considers an access point (AP) connected to multiple IoT devices assisted by RISs and acts as an FL server, while the RIS and IoT devices are considered clients. The FL model is trained based on location information and optimal RIS configuration. As a result, the trained FL model can achieve high transmission sum rates in IoT networks.

*5) IoT Networks:*

High-dimensional data analytics will shift the traditional IoT paradigms from connected things to connected intelligence. It is envisaged that FL will be an indispensable tool in intelligent IoT-based applications, which are spreading in diverse fields [80], [81]. In this section, we outline the usage of FL in various sectors associated with IoT networks.

***Industrial IoT (IIoT):*** The fourth industrial revolution (Industry 4.0) was triggered by the advancements in automation and manufacturing industries, coupled with the emergence of IIoT devices. Albeit the promising features of FL can be beneficial for IIoT networks, the upsurge number of nodes that may participate in the training process might produce colossal traffic that burdens the network. Reliable participant selection schemes can reduce network overhead and alleviate communication costs. The work in [82] presents a budgeted client selection algorithm that enhances the global model accuracy by choosing the best clients. This algorithm finds $R$ clients with the best test accuracy based on the secretary problem. More specifically, clients are interviewed sequentially and marked as selected or rejected, and then these clients will be ranked from the best to the worst to facilitate the selection process. Another serious design aspect in FL-empowered IIoT networks is edge device failure, which causes severe fluctuations in production quality. The authors in [83] shed light on such aspects and propose an anomaly detection framework that uses FL to train edge devices to predict abnormalities, enabling enhanced communication efficiency.

***Healthcare applications:*** FL has become very popular in the field of healthcare applications [84]. Pandemics negatively affect human health and cause negative impacts on economics. Recently, Covid-19 swept the world, causing health problems and mortality. Covid-19's primary manifestation is pneumonia which is detected using X-ray scanning. ML can play a vital role in such medical cases, in which collected data can be exploited to train an ML model that can predict the infectious state. By emphasising that patient data across different medical centres should be handled privately, the FL setting is the natural option for such applications. Therefore, Liu *et al.* [85] applied FL to datasets of various clinical centres; FL clients exploited the available local X-ray images of Covid-19 cases at each hospital to train a model that helps practitioners to determine if a patient has been infected, without leaking any personal information.

***Financial Perspective:*** The financial sector plays a central role in all societies. In particular, the dependency on credit cards has exponentially increased in recent years, facilitating everyday life. Security attacks constitute a major threat to credit card systems, resulting in critical information leakage and money loss. Currently, banks utilise their datasets individually to develop centralised ML algorithms for fraud detection to mitigate such threats, but this was unavailing as the datasets did not help create an accurate model due to their insufficiency. To overcome this challenge, the work in [86] presented a framework that depends on FL to build a fraud detection system that is collaboratively trained using datasets from multiple banks. The problem is that the number of fraudulent transactions is too small compared with legitimate transactions; this can obstruct FL performance. To this end, the synthetic minority over-sampling technique (SMOTE) is used to oversample the minority class by producing synthetic datasets that can be used to train the FL model for enhanced model inference.

### *B. FL Potential Future Applications*

After describing FL and presenting its applications in various wireless networks, we outline some prospective application scenarios in new and promising areas. Our vision is primarily inspired by the applications anticipated to be inherent in B5G and 6G networks.

#### *1) Visible Light Communications (VLCs):*

VLC is a new nascent wireless communication technology that relies on the visible spectrum for data transmission. VLC exploits the advantageous properties of light-emitting diodes (LEDs), such as low-power consumption, high brightness, and a long lifetime, to provide high data rate, low latency, and green indoor communications [87]. VLC will play a major role in relieving the pressure on the scarce spectrum of the current wireless networks and provide a new connectivity method for the ever-increasing IoT devices. Fig. 5 represents the VLC communication system that consists of LED units, called APs, connected to a gateway that, in turn, is connected to the external network through wired or wireless links. As a subfield of AI, FL will have a role to play in promoting VLC applications. The features that characterise VLC systems,

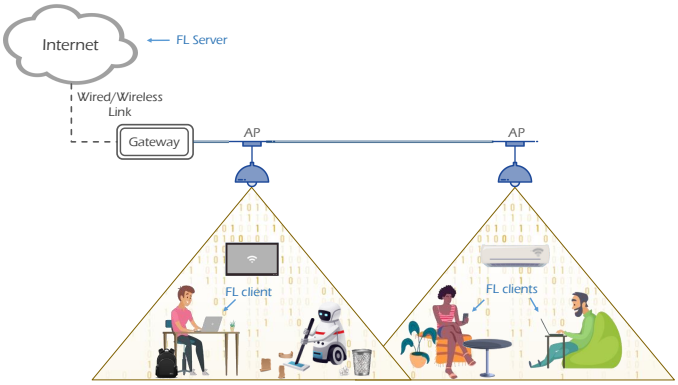


Figure 5: VLC system uses visible light as a medium for communication for wireless devices.

including high spatial reuse, ultra-low-latency, ultra-high-data rates, and inherent security, provide the ingredients needed for the efficient implementation of FL algorithms [88]. The main purpose of FL is to secure data privacy and reduce communication overhead. Accordingly, in the VLC network, the external server can be the FL server, and the deployed indoor devices can play the role of FL clients. In this case, clients may leverage the fast, secure, and reliable transmission environment to update the global model through the gateway, while the FL server can reach the required level of convergence faster.

FL training latency highly depends on client selection and scheduling. Therefore, how to properly select FL clients in the VLC network is an important question that needs to be addressed. Besides, as the number of participants increases, the global model can better infer accuracies. Nevertheless, the field of view of the LED units is limited and covers a limited number of clients so that a few devices can participate in the FL training process. To increase the number of participants in the VLC network, the HFL can be utilised, where the APs are used to aggregate the model updates of the clients under their coverage. Once this step is completed, the APs send the aggregated models to the central FL server. One interesting application of FL in VLC is predicting when an LED will stop illuminating due to, for example, LED life expiration or LED light-off time and instructing the endpoints to an alternate connection. Additionally, FL can play a major role in predicting clients' mobility, LED beam assignment, and client-LED association, to name a few.

#### *2) Cell-Free Massive MIMO (CFmMIMO):*

The implementation of massive multiple-input multiple-output (mMIMO) networks includes two types based on the antenna deployment strategy: collocated and distributed antenna setup. The collocated type is easier to implement and has low data sharing overhead, which requires less backhaul. In contrast, the distributed implementation is more complex but gives the network an improved performance, especially in coverage gain. Recently, a new promising technology called CFmMIMO has been proposed as an incarnation of the distributed antenna setup [89]. CFmMIMO constitutes a radical change in the cellular network paradigm as it eliminates

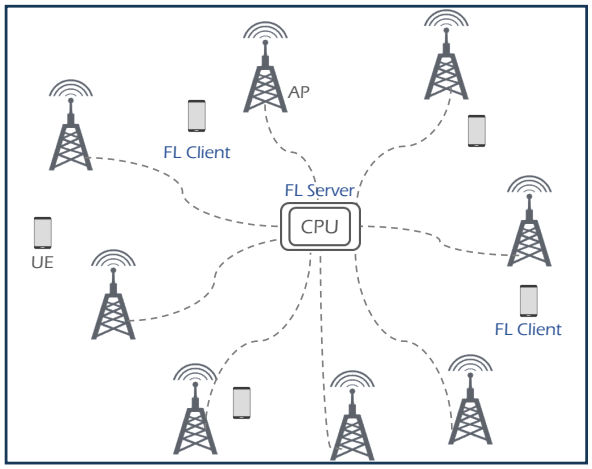


Figure 6: Cell-free mMIMO network which shows number of UEs surrounded by many APs.

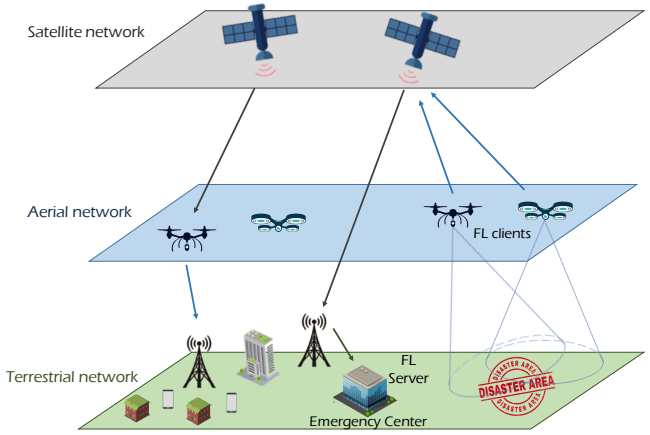


Figure 7: Illustration of satellite-aerial-terrestrial integrated networks. Satellites are used as a relay for communication between UAVs and terrestrial BS.

the concept of cells. Many small, simple, geographically distributed BSs, called APs, jointly serve a small number of UEs using the same time-frequency resources via time division duplexing. APs are connected to a CPU through backhaul links and use the fronthaul to serve the UEs simultaneously, as shown in Fig. 6. CFmMIMO enhances the UEs connectivity by eliminating inter-cell interference and reducing the path attenuation due to the presence of the UEs near the APs.

CFmMIMO embraces distinct features that have a significant advantage in favour of FL. One such features is channel hardening [90], which means that the fading channel will behave as an almost deterministic scalar channel. Channel hardening greatly benefits FL, especially when selecting the clients to participate in the training process. Selecting UEs with stable connections eliminates any unfavourable transmission failure when uploading local updates to the FL server, i.e., CPU, thereby enhancing the FL performance. Moreover, when many APs surround the UEs, this will lead to high coverage gain and reduced distance between the UE and the AP. As a result, this will facilitate training the global model that requires a large number of clients to participate in the training process, thus reducing training latency and improving performance. Furthermore, FL can realise potential applications in CFmMIMO, for instance, creating FL models capable of assigning users to the optimal APs that fulfil the desired QoS by measuring the received signal strength of many surrounding APs. On the other hand, FL can be used to alleviate the congestion on the APs by training a model that can monitor, predict, and distribute UEs to APs in a way that maintains network performance.

*3) Satellite-Aerial-Terrestrial Networks:*

Terrestrial cellular networks aim to serve populated regions while building such networks to serve sparsely populated areas like islands, oceans, and mountains is impractical. Satellite communication systems address this issue by providing rural areas with network connectivity. However, the quality of satellite links is not guaranteed due to challenges such as large path loss and limited UE power transmission. For this reason, the research has been directed toward utilising aerial platforms to aid satellite communications. High altitude platforms (HAPs) can be used to provide broadband services over a large coverage area [91]. Moreover, HAPs provide more reliable communication links than terrestrial networks because they are less susceptible to ground blockages and multipath signal effects. Integrating the aerial-satellite network forms a space-backbone network layer that can provide wireless connectivity to ground users anywhere. As a result, the hybrid satellite-aerial-terrestrial networks have drawn the research community's attention for further improvements, which are envisioned to be an essential part of the B5G/6G networks. Fig. 7 represents the topology of the satellite-aerial-terrestrial network.

Recently, ML techniques have been considered in solving challenges related to satellite communications [92]. Employing FL in satellite-aerial-terrestrial networks is still in its infancy; thus, there is plenty of room to explore the potential of FL in such networks. For instance, FL can tackle the network's limited resources, security, and energy usage challenges. Furthermore, satellite and aerial platforms have received significant attention due to their ability to deliver services in emergency scenarios such as disaster relief, and rescue missions [93]. To achieve this, it is necessary to maintain robust and reliable communication between satellite, aerial, and ground-based networks. For instance, the terrestrial networks may be overloaded or destroyed if a large-scale disaster occurs, demanding a rapid establishment of a network to serve the afflicted area. Airborne vehicles can cover and monitor this area and send information to an emergency centre. However, in some cases, the vehicles may be outside the coverage of terrestrial BS; therefore, the vehicles can establish a connection with a satellite to act as a relay point between the air vehicle and the terrestrial BS, as demonstrated in Fig.7. Airborne vehicles provide the needed multimodal information; however, transferring a large amount of data burdens the communication links and consumes much time, which is critical in such situations. In this case, employing FL can eliminate the drawbacks. Equipping the vehicles with a pre-trained FL object detection and localisation model allows for sending lower-size vital information to locate survivors.

Simultaneously, the vehicles can train the model using the collected data to enhance its accuracy and then send the model updates to the FL server. Accordingly, saving time and relieves communication links.

*4) Semantic Communication:*

The main theme of communication systems up to 5G networks was to ensure the correct reception of every single transmitted bit, regardless of the meaning conveyed by the transmitted bits. However, this classical communication-theoretic framework does not meet the aspiration of B5G/6G networks, as the research community agrees on the need to upgrade this framework to a smarter and more informative one. The overlooked meaning behind transmitted data is expected to play a significant role in next-generation communication systems, forming an interface between machine intelligence and human intelligence. Therefore, considering data content's high-level meaning or relevance to support machine-intelligent services necessitates a shift from semantic-neural toward semantic communication systems [94].

The interplay between human beings and AI has resulted in many revolutionary applications like virtual reality (VR)/augmented reality (AR) and haptic communications. Several studies have begun to envision the integration of FL with semantic communications, VR/AR, and haptic communications. Similarly, in this article, we discuss several wireless scenarios where FL can be applied in these emerging fields. In semantic communications, using FL helps improve the network's BW utilisation by training a model that can extract relevant/contextual information from the data and filter out irrelevant information. FL-based semantic communication can effectively preserve the network's resources by transmitting semantic information rather than bits or symbols. On the other hand, VR/AR are real-time technologies that bridge the real and digital worlds by replacing or enhancing the physical environment with a computer-generated one. In the AR/VR environment, detecting users' movement and location is essential and heavily influences the wireless network's resources. FL is effective in predicting user movement and actions, which can be used to optimise the allocation of wireless resources to users [95]. Finally, haptic communications bring a new dimension over conventional communication modalities by enabling real-time haptic experiences between tactile parties. Haptic communication will have diverse applications, particularly in industry and health sectors, which poses a critical need to protect such communications. FL is a vital tool for securing haptic-based applications through training an ML model that can discriminate between genuine and counterfeit actions based on previous signatures and warn the system of possible suspicious measures.

## V. FL CHALLENGES

Deploying FL in various fields demonstrates its efficiency and highlights its main advantages. However, the successful implementation of FL is restricted by some challenges and limitations that must be resolved to realise its full potential. In this section, we articulate the most common challenges of FL and outline their proposed solutions. Table II summarises the key FL challenges and the associated solutions.

### A. Server Centralisation

The performance of employing FL depends by large on the server and the participants. The bottleneck of either classical FL or HFL systems relies on the dependency on a centralised server to orchestrate the learning process, representing a single point of failure. Additionally, the large number of model updates sent to the central server can overwhelm the network, resulting in traffic congestion and degrading the network performance. Two approaches were used to address this challenge, namely blockchain and peer-to-peer.

***Blockchain approach:*** Adopting FL systems integrated with blockchain instead of a central server avoids malfunctions that may result from using a single centralised server. Blockchain has been widely used in the literature [62], [68], where it can provide a distributed, end-to-end trustworthy training environment. The blockchain consists of miners and devices; miners can be randomly selected devices or separate nodes (such as cellular BSs or WiFi APs) that are computationally powerful to perform the mining process. The operation of blockchain-based FL systems can be summarised as follows: the process begins at the participating devices by computing and sending the local model updates to the associated miner in the blockchain network. Next, miners verify and exchange the local model updates using one of the consensus algorithms, generating a new block where the verified updates are recorded. Finally, the generated blocks that store the model updates are added to the blockchain and can be downloaded by the devices to perform the next round of computation. Leveraging the blockchain will not adversely impact the overall network system when a failure or malfunction happens in a miner, making the FL system more robust.

***Peer-to-Peer approach:*** The study in [96] proposed a new technique called BrainTorrent, in which a centralised server is not required. This technique is aimed at medical applications where data sharing is prohibited due to privacy concerns. According to the authors in [96], BrainTorrent is a peer-to-peer procedure where each centre shares its model updates directly with the others without needing a central body to coordinate the process. Initially, every client maintains a version of the trained and old models. One of the clients in the network initiates the training process by sending ping requests to all other clients to update the model. Other clients will respond by sending their model weights and the training sample size. Then, the model weights are aggregated and averaged at the request initiator based on the clients' dataset size to produce a new version of the trained model, followed by repeating the process until a certain level of accuracy is attained. The main drawback of this technique is that it is feasible for networks containing a limited number of clients, while in an environment with a large set of clients, such a technique is impractical.

### B. Clients Selection

To recall, the FL process consists of three main phases: selection, configuration, and reporting [30]. These three phases are performed iteratively until the FL model achieves a satisfying level of model accuracy. In the selection phase, the

Table II: Summary of FL challenges, impacts, and proposed solutions.

| Area | FL Challenge | Impact | Solution | Article |
|---|---|---|---|---|
| Cellular Networks | Client Heterogeneity | Synchronisation Issue | Asynchronous FL Scheme | [46] |
| | Securing Model Updates | Private information leakage | Blockchain Technique | [46] |
| | Server- Clients Communication Latency | Reduced Model Aggregation Efficiency | HFL | [27] |
| | | | Resource Allocation Technique | [55] |
| | Training Accuracy | Reduced Model Accuracy | Participants Selection Scheme Based on the Weights | [55] |
| IoV Networks | Server Centralisation | Single Point of Failure | Blockchain Technique | [62] |
| | | | Asynchronous Peer-to-Peer Updates | [97] |
| | Data Heterogeneity | Degraded Model Accuracy | Model-Selective Approach | [94] |
| | Client Mobility | Server-Client Communication Disruption | Asynchronous FL | [97] |
| | Hyperparameters Selection | Inefficient Model Training | PSO Local Parameters Tuning | [66] |
| | Client Incentivisation | Reduced Model Accuracy and Longtime Training | Blockchain Loyalty Program | [62] |
| | Client Heterogeneity | Reduced Model Aggregation Efficiency | Model-Selective Approach | [64] |
| | Securing Model Updates | Private Information Leakage | DP Technique | [97] |
| UAV Networks | Server Centralisation | Single Point of Failure | Blockchain Technique | [68] |
| | Securing Model Updates | Private Information Leakage | Local DP | [68] |
| | Client Incentivisation | Reduced Model Accuracy and Longtime Training | Two-Tier RL- Based Incentive Mechanism | [68] |
| | Data Heterogeneity | Degrade Model Accuracy | Dempster Shafer Client Grouping Based on Data Quality | [74] |
| | Server- Clients Communication Latency | Reduced Model Aggregation Efficiency | Adjusting the CPU-Frequency and Upload BW | [70] |
| | Clients Energy Consumption | Increased System Cost | Adjusting the CPU-Frequency and Upload BW | [70] |
| RIS-Assisted Networks | Server- Clients Communication Cost | Reduced Model Aggregation Efficiency | AirComp Technique | [75] |
| | Client Heterogeneity | Reduced Model Aggregation Efficiency | Unified Communication-Learning Optimisation | [76] |
| | Unreliable Wireless Channels | High Transmission Delay | RIS-Receiver Joint Optimisation | [77] |
| | Server-Client Transmission Throughput | Reduced SINR | RIS-based Optimisation Scheme | [78] |
| IoT NETWORKS | Client Heterogeneity | Synchronisation Issue | Personalised FL | [99] |
| | Data Heterogeneity | Degrade Model Accuracy | Personalised FL | [99] |
| | Model Heterogeneity | Different Model Architecture | Personalised FL | [99] |
| | Server- Clients Communication Cost | Reduced Model Aggregation Efficiency | Budgeted Clients' Selection | [82] |
| | Client Abnormality | Degrade Model Accuracy | Anomaly Detection Framework | [83] |
| | Unlabeled Data | Add Complexity to Model Training | Pseudo Labeling Technique | [98] |
| | Data Insufficiency | Reduced Model Accuracy | Oversampling and Producing New Synthetic Data using SMOTE | [86] |

server determines the optimum users allowed to participate in the training process according to predefined selection criteria, i.e. whether or not the device is available and its resources. Concerning clients, two main factors directly impact the model convergence speed and efficiency.

*1)* ***Clients heterogeneity****:* In practical wireless networks, end devices have different hardware characteristics and experience varying channel transmission conditions, in addition to data heterogeneity. For instance, clients with high hardware capabilities and high-quality data can produce a well-trained model in a relatively short period compared to others. Furthermore, clients experiencing good transmission circumstances support low latency model transmission, enabling timely parameter aggregation. Failing to consider these aspects will reduce the efficiency of the FL training process. In [97], a participant selection scheme based on the available client resources has been proposed. Rather than selecting random clients, this scheme sends a resource request to the clients to collect information about their hardware specifications, communication reliability, and data availability. Based on this information, the server estimates the time required to complete a specific task and then selects clients that will participate in the following training round accordingly. In a similar context, the scheme in [64] relies on selecting local models based on the clients' computation capability and data quality.

*2)* ***Clients incentivisation****:* FL depends on the participants and the on-device datasets. The high computation resources and valuable data attract FL to select these clients for training.

However, nothing forces the end device to participate in a learning process that will deplete its resources and leads to unsolicited costs. Thus, a reward procedure must be considered to encourage the end devices to participate in the FL training process. In this context, various client incentivisation schemes are released [62], [68], [98]. In [62], a loyalty program based on the blockchain technique is presented to motivate users with large samples of useful data to participate in FL training. According to their contribution, the loyalty program rewards the participants, attracting users with high data quality to participate in the training process.

Furthermore, a two-tier RL-based incentive mechanism is presented in [68]. The two-tier RL mechanism enables obtaining the best scenarios for the task publisher and clients in a dynamic environment by encouraging the workers to provide high-quality model training when explicit network parameters are unavailable. The reward of each client is maximised based on the contribution provided to enhance the global model. The work in [99] improves the reliability of FL by proposing an incentivisation scheme that combines the client reputation and a contract theory to encourage clients with high-quality data to participate in model learning.

### C. Data Heterogeniety

Data is the main driver of ML algorithms, and high-accuracy model training requires a large amount of data. Generally, practical datasets are heterogeneous and require pre-processing before they can be used for model training.

***Data quality:*** The characteristics of the locally generated data differ from one user to another. Datasets can be classified into two categories, independent and identically distributed (IID) and non-IID data. In practical scenarios, datasets are usually non-IID [100], while most of the existing literature in FL is based on the assumption of IID data. Given data heterogeneity, Ye *et al.* [64] employ a selective model aggregation approach to evaluate the quality of the images and then quantify it. The central server evaluates the image quality based on the clients' historical records and prepares a contract to select fine clients with fine models. In [74] Dempster_Shafer technique is used at the global node to classify and prioritise the UAV clients into groups according to data quality. The highest priority group can contribute more to model training and produce better model weights.

Fairness in FL has recently received more attention. As data heterogeneity increases among clients, the training process will produce a skewed model that may ignore some of the clients, resulting in fairness issues. A possible solution is to employ the personalisation concept [101], where the global model with coarse-grained features is sent to each participating device, and then the clients train the model using their data to build a model with fine-grained features.

***Data insufficiency:*** In some cases, the data collected by the devices may not be large enough to conduct model training. On the other hand, the percentage of high-quality data could be small compared to the total datasets, which affects the model inferencing and classification tasks. The proposed work in [86] uses the SMOTE technique, which attempts to rebalance the classes in the datasets by oversampling the required features' data. SMOTE generates new synthesised data examples close to the observed datasets. Another approach that can be used to provide more data samples is based on generative adversarial networks (GANs) [102]. The goal of GAN models is to study and determine the distribution of the training data samples to generate more close to actual data samples from the estimated distribution.

***Data annotation:*** Most studies considering FL assume a supervised training approach, where the data is processed and classified to facilitate the training process. However, in real situations, most of the generated data is unlabelled; in this case, the unsupervised FL is the method that should be considered. Data annotation is a challenging task that requires high cost and significant effort. The presented work in [103] uses a pseudo-labelling technique to classify the unlabeled data based on the labelled data. Instead of manually labelling, which is time-consuming and requires much cost and effort, pseudo-labelling gives approximate labels depending on the model trained by the labelled data. The FL algorithm is used to train the model in two phases. First, the global model is trained by the distributed devices' labelled data until reaching a certain convergence level. Second, improving the performance of the trained global model by training it again using the classified unlabeled data.

### D. Communication Cost

The model convergence speed and accuracy in FL highly depend on the hardware specifications of the server and the clients. Despite the recent advancements in the computational and communication capabilities of end devices, model training and transmission overhead over multiple training rounds remain major design issue that potentially affects the global model training quality. Furthermore, a large number of model updates exchanged between the server and the clients can severely exhaust the network communication resources. In the following, we outline the main approaches to tackling this challenge.

***Models scaling and superposition:*** Despite the significant advancements in edge computing, the lack of communication resources in current FL-enabled systems seriously affects the latency performance and reduces the model convergence rate. To this end, AirComp has been proposed to provide a co-design approach for the FL aggregation procedure by utilising the superposition nature of radio channels for simultaneously transmitting model updates from different clients [104]. Therefore, improving communication efficiency by reducing the required BW resources and providing fast convergence. Later, a new variant of AirComp was introduced, called broadband analog aggregation (BAA) [105] to cover wideband channels that can carry the multidimensional updates of local models. Furthermore, in [106], the authors propose a framework for model aggregation that relies on digital modulation. The proposed scheme utilises a single-bit gradient quantisation and quadrature amplitude modulation at the edge devices to achieve fast model convergence.

***Resource allocation**:* It was demonstrated in [70] that joint optimisation of onboard computation resources and BW allocation can be a promising solution to the computation/communication overhead in resource-constrained devices. This issue is more pronounced in ultra-dense networks, where an excessive number of model updates must be exchanged for global model convergence, and this further yields traffic congestion. Thus, selecting a subset of clients has proved its efficiency in tackling the communication cost problem. In particular, employing only clients with high-quality data in the training process can speed up the convergence rate. Subsequently, a reduced number of training rounds will be performed. Yao *et al.* [107] propose a two-stream model approach to reduce the FL communication cost. In this approach, the single model that was typically used to be trained by the clients is replaced by a two-stream model. The authors exploit the transfer learning mechanism and maximum mean discrepancy to force nodes to learn other nodes' knowledge. The experimental results showed a reduction in the required communication rounds and reduced communication costs.

***Gradients compression**:* Large-scale deployment of FL requires significant communication rounds between the central server and the clients. This necessitates expensive network resources to perform model parameter exchange, which can limit the scalability of the FL system. The work in [108] significantly reduces communication costs by using deep gradient compression. The work shows that most SGD parameters are redundant, so the compression technique is employed to considerably reduce the number of transmitted parameters while preserving the model's accuracy. Moreover, the compression minimises the gradients by sending only the necessary gradients to the central server in each round, reducing communication latency and alleviating the utilisation of limited wireless resources. On the other hand, by exploiting non-orthogonal multiple access (NOMA), a new 5G medium access technology that improves spectrum efficiency by allowing simultaneous transmission over the same channel, the work in [109] proposes a NOMA-enabled adaptive gradient compression FL system. In this work, the authors exploit NOMA and adaptive gradient quantisation and sparsification to facilitate uploading model updates over fading wireless channels.

### *E. FL Latency and Convergence*

The network and devices heterogeneity, data statistics heterogeneity, dynamic wireless environment, and acquiring the CSI are the most important factors influencing FL performance in terms of latency and convergence rate. An appropriate client scheduling mechanism can be the key to an accurate and fast model convergence. The authors in [110] formulate an optimisation problem that jointly selects a group of clients with local models that significantly impact the global model and assigns the limited resource blocks to those clients. Furthermore, Huang *et al.* [111] proposed a stochastic client selection algorithm that jointly considers the cumulative effect of participants and selection fairness to maintain a high-quality training performance while ensuring fairness among high-qualified and low-qualified clients. Moreover, enabling edge computing can remarkably reduce the FL latency, in which APs are placed close to the edge device, and hence, reduced latency can be achieved. Within the same context, the authors in [112] proposed a framework to reduce the average time per round by considering latency-based scheduling, in which clients are selected based on their computation and communication delay.

Generally, ML algorithms are sensitive to hyperparameters, which play a critical role in the model convergence rate. Therefore, to further reduce FL training latency and enhance the convergence time, careful consideration should be taken in the design of efficient hyperparameters. In this regard, several hyperparameter tuning algorithms have been developed to manage many of these parameters with their wide ranges. This includes Bayesian optimisation, grid search, and random search. The work in [113] develops a scheme that can efficiently determine the optimum learning rate (LR) values. In the proposed technique, referred as cyclical LR (CLR), the CLR is bounded by a range of carefully selected values, in which its value can vary. This approach aims to avoid random LR initialisation. The presented results in [113] showed the efficiency of such a technique in reducing the FL latency by minimising the number of training operations while ensuring a particular level of accuracy.

### *F. Securing Model Updates*

Although FL is motivated by inherent privacy-preserving and security features, sophisticated intruders can retrieve critical information about the participating nodes from the shared model updates. Besides, malicious devices may opt to participate in the training model process to inject false model updates, affecting the accuracy of the trained model. The following approaches are developed in order to ensure secure model transmission:

*1)* ***Secure multi-party computation (SMC)**:* A cryptographic protocol that aims to conceal personal information and guarantee zero-knowledge between multiple involved parties [114]. Its main idea is to distribute the computation between multiple parties without exposing or moving private information. Its working mechanism can be summarised as follows: first, the participated organisations' datasets are split and masked by adding random numbers, and then these encoded segments are shared between organisations to perform the required computation, thus guaranteeing data privacy and trust. SMC allows organisations to work together without knowing one anothers' confidential information.

*2)* ***Differential privacy (DP)**:* This approach prevents leaking model parameters to intruders by leveraging artificial noise, which is added to the locally trained model before transmission [115]. However, enhanced security comes at the expense of model accuracy; hence, joint optimisation is essential to strike a balance between security and model accuracy. Such technique has been used in the literature, e.g., [116] and [68], in which random Gaussian noise is utilised to enhance the privacy of model parameters.

*3) **Homomorphic encryption (HE)**:* HE is a key-based security mechanism which allows performing calculations on encrypted data. In the context of FL, participating clients generate public and private keys, where the former is used to encrypt locally trained models. After that, the model updates received from all clients are aggregated on the server side in an encrypted mode. The clients leverage the private keys in order to decrypt the global model updates. Albeit the enhanced security achieved by exploiting the HE mechanism, the computation complexity of the cryptographic operations imposes additional overhead on the resource-constrained clients in terms of time, power consumption, and communication cost. In this regard, Zhang *et al.* [117] proposed a batch encryption technique, which minimises the encryption and communication cost resulting when using HE. Specifically, each client quantises the gradients to be represented in a low-bit integer format, and then a batch of the encoded gradients is encrypted for transmission. Consequently, the encryption overhead and the size of the total ciphertext will be considerably decreased.

## VI. FUTURE RESEARCH DIRECTIONS

Despite the prospects brought by the advancements of FL, its application is still in its early stages. This necessitates dedicating the research efforts toward addressing the associated challenges and exploring new horizons of implementation possibilities. In the following, we list a number of interesting future research directions.

### A. Data Freshness

In information technology, data is marked by the date of its creation and can become meaningless, i.e., outdated. Access to timely information (i.e., data freshness) is paramount for time-based systems driven by datasets [118]. In order to quantify the data freshness, the age of information (AoI) metric is introduced [119], and is considered an essential parameter in realistic scenarios of data networks. From the perspective of FL, AoI can be defined as the time that elapses between collecting data from clients and completing the FL training task. Considering applications with tight latency and throughput requirements, e.g., ITS, the AoI becomes crucial in network design principles. Accordingly, future research may focus on proposing novel schemes that select FL clients based on their data freshness to ensure that the required network reliability is achieved. Additionally, distributed client datasets can be highly temporal and change rapidly; thus, incorporating the rapidly changing data and determining the correct timing of model updates is essential to enhance FL performance in highly dynamic environments.

### B. Spectrum Sharing

The widespread use of IoT devices and the new technological trends make the limited spectrum bands insufficient to meet the requirements of BW-hungry applications. To this end, spectrum sharing is proposed to mitigate the pressure on frequency bands by allowing multiple networks to operate using the same portions of the licenced or unlicensed spectrum, provided that they do not interfere with each other [120]. Coexisting networks should consider interference problems, i.e., co-channel and adjacent channel interference, addressed by imposing strict rules from telecom regulators. Multiple networks from the same or different technologies can coexist and use the same spectrum band, where this coexistence is categorised into equal and different access rights. The major concerns associated with equal rights coexisted networks are maintaining seamless operation, mitigating harmful interference between them, and ensuring fairness. By exploring the literature, we conclude that it is difficult to satisfy these concerns without the intervention of a third party who must receive information from the coexisting networks and manage transmissions. However, this method is undesirable as it requires information disclosure and incurs additional communication costs. Therefore, the FL algorithm is a potential solution that preserves network data privacy and eliminates the need for a third party. Coexisted networks transmission demands and local spectrum utilisation can collaboratively train a global FL model, for instance, deep RL, to address coexisting issues. This model is fed back to each network to make the right spectrum access decisions.

### C. FL at Scale

The applications mentioned in Section IV-A are considered small-scale scenarios. However, many applications require a wide deployment of FL to take advantage of the data collected in different locations. This helps to get feature-rich datasets from extensive scenarios that can train an effective global model. Designing an FL system that covers large-scale environments requires special attention to the FL server capabilities in addition to cellular and backhaul communications. The number of participants can easily reach millions spread in broad areas and produce massive model updates that must be transmitted through the wireless network. Therefore, considering the network's communication efficiency alongside selecting an FL server with efficient hardware to handle enormous amounts of updates is crucial. To this end, future research should consider wireless network design and the specifications of the FL server suitable for large-scale deployments and develop a technique that intelligently selects the optimum participants among many devices willing to participate promptly.

### D. Meta-Learning

The shortcoming of existing ML techniques, especially DL algorithms, is that they rely on large datasets to develop a good model. In most cases, it is not possible to obtain a high amount of dataset, while in other cases, the number of samples that hold the desired features is small compared to the entire dataset. Therefore, finding a mechanism to train models based only on a small dataset sample is necessary. In light of the preceding discussion, the meta-learning technique is introduced to address data insufficiency [121]. Meta-learning, also known as learn to learn, uses the metadata of other tasks, like data patterns, properties of the learning problem, and the algorithm performance to learn how to learn and then learn the new task more efficiently from a small set of data. This new

learning method in the FL algorithm is expected to improve its performance in several aspects. First, using only a small amount of data samples in meta-learning leads to a more convenient client selection. Moreover, the operating cost will be reduced, thus saving many resources and training time. The optimal client selection will also lead to rapid model convergence and lower latency which is crucial for B5G and 6G networks. Finally, meta-learning can help adapt the global model to each user, especially when data heterogeneity exists among clients.

### *E. Modality Agnostic Learning*

In current ML approaches, the models are designed based on the characteristics of input data dedicated to a specific task. However, B5G/6G networks allow the creation of different dataset modalities, such as vision, audio, time series, and point cloud. When a specific model needs to be used with a different data configuration, its architecture must be redesigned. This means that best-practice models cannot be used in different domains without modification. Perceiver [122] is an interesting solution proposed to handle the configuration of different data shapes based on Transformers networks [123], which are sequence transduction models that rely entirely on the attention mechanism. The usage of Transformers in computer vision has shown their efficiency in classification tasks using considerably lower computation resources. Therefore, utilising dynamic models that suit multimodal inputs, like Perceiver, in the FL setting will help in its realisation and wide adoption to perform different network optimisation tasks that render the network more reliable. In more detail, dynamic models will allow clients with different data shapes to participate in the FL process, facilitating FL operation. This new research direction needs further investigation under the umbrella of FL systems.

### *F. FL Carbon Footprint*

DL-based approaches are highly dependent on heavy computations, resulting in high power consumption. Higher energy cost increases carbon dioxide equivalent ($CO_2$e) emissions, constituting the main reason for climate change [124]. Recent studies have been devoted to investigating the impact of ML on Earth's climate, steering the focus to the environmental effects of training large-scale ML models connected to network grids powered using fossil fuels. The environmental consequences of FL in wireless networks have not been explored much; few studies have recently begun to detail such implications. In addition, the transition from centralised to distributed learning seems more energy efficient. Avoiding transmitting big data to a central location saves much network energy and eliminates the need for cooling and other auxiliary tasks. However, the ML technique and the number of participants determine how efficient the network is. With this in mind, the study in [125] proposes a sustainable FL-based framework by considering energy harvesting technology. Our vision is that future wireless networks will highly depend on renewable energy resources; for instance, we may see more dependence on solar power at the edge devices. This aspect opens the horizons for exploring FL approaches that can potentially contribute to achieving carbon-friendly wireless networks. To this end, future research should be dedicated to assessing the environmental impacts of FL-empowered networks before being widely used in broader scopes.

### *G. Low-Precision FL*

Computational capabilities are a significant factor in determining the best clients involved in the FL process. However, edge devices often have limited computing resources, making implementing FL more complex. For instance, the computational complexity of DL models increases as the model becomes deeper, requiring high-performance hardware, while in reality, resource-limited devices are available. The use of full-precision DL models that perform floating-point mathematical operations is a major reason for the increased computational complexity of such models. Various approaches are introduced to compress deep networks, such as parameter pruning [126] and parameter quantisation [127]. Much interest has focused on the model quantisation technique as it produces more compact models than their floating-point counterparts. Binary neural networks (BNN) [128] is a promising approach that recently emerged to facilitate deploying DL models in resource-limited devices. In BNNs, model weights are quantised using binary values. The merits of BNNs represented in memory saving, computation reduction, and energy efficiency make them appealing for use under the FL setting. The combination of FL and BNNs will form a new low-precision framework that can be used at the edge of wireless networks. Although the usage of BNNs addresses the scalability of the FL process, their performance is degraded compared to other full-precision counterparts. Using BNNs in FL is a promising solution; nonetheless, more research should focus on optimising BNN-based FL frameworks and closing the performance gap.

### *H. Digital Twin (DT)*

DT is a technology representing a physical object, service, or even an entire system in its counterpart digital version [129]. The DT framework aids the operation of complex systems by providing insights into how these assets behave under various simulated circumstances that will help improve decision-making and optimise these systems. As reported by Gartner, DT is envisioned to be one of the most influential industry 4.0 technologies in the next decade. Furthermore, DT is a data-driven technology that can provide system operation excellency by leveraging real-time analysis when paired with AI. However, The DT faces the challenges associated with big data and privacy protection. Accordingly, a novel collaborative paradigm can be achieved when fusing FL with DT systems to meet these challenges. As two emerging and promising techniques, FL and DT can help reduce wireless networks' operation complexity and realise 6G-based IoE applications [130]. Despite the literature's scarcity of works that leverage such fusion, it is envisioned to become an essential part of the next generations of wireless networks. Future research may consider using FL with DT to share knowledge between DT nodes and develop a common understanding. In addition, the DT can assist FL tasks, for example, by quantifying the DT node's trust and selecting clients based on the degree of trust.

### I. FL Task-Reward Announcement

FL model training depends on the participating clients' resources and the corresponding on-device datasets. Data quality differs from client to client based on usage, and behaviour [131]. The selection of devices is based on predefined conditions like being connected to an unmetered network, idle, and in a charging state. Moreover, choosing the optimal clients for a particular task helps relieve the pressure on wireless spectrum resources by lowering the training rounds required in FL and improving network latency. However, to encourage users to participate in the FL process, a reward mechanism should be developed to compensate for their consumed resources and data used while training. How to determine and select participants based on their resources and data quality is ongoing research. FL task-reward announcement is a necessary approach. With an effective announcement technique, users with high-quality data and resources may be encouraged to make themselves ready to participate in the FL process by matching the terms of participation required to receive some rewards. Announcement techniques can enhance the overall performance of the FL system.

## VII. CONCLUSIONS

The emergence of FL and its distinctive features pave the way for numerous advancements in the industry. Motivated by the various implementation scenarios in different wireless networks, we conducted a survey demonstrating the salient merits of FL. In this context, this review paper presented the basic operational principles of FL and discussed the essential enabling technologies. This is followed by a discussion of state-of-the-art wireless network applications optimised by utilising the FL mechanism. Moreover, we shed light on promising research directions that may unlock the potential of FL in new areas of B5G and 6G wireless communication systems. Furthermore, we focused on the challenges associated with implementing FL and outlined the techniques used to address those challenges in literature, and then we offered insights to improve the design of the FL algorithm. We believe that the way this survey is harmonised can offer a firm understanding of FL usage in various areas, facilitating the focus on new research directions.

## ACKNOWLEDGEMENT

This article is supported by Ajman University Internal Research Grant No. 2022-IRG-ENIT-18. The research findings presented in this article are solely the author(s) responsibility.

## REFERENCES

[1] D. Gil *et al.*, "Internet of things: A review of surveys based on context aware intelligent services," *Sensors*, vol. 16, no. 7, p. 1069, July 2016.
[2] Z. Rehena, "Internet of Things," *Interoperability IoT Smart Syst.*, p. 1, Dec. 2020.
[3] M. Shafi *et al.*, "5G: A tutorial overview of standards, trials, challenges, deployment, and practice," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 6, pp. 1201–1221, Apr. 2017.
[4] H. N. Dai *et al.*, "Big data analytics for large scale wireless networks: Challenges and opportunities," *ACM Comput. Surv.*, vol. 52, no. 5, pp. 1–36, Sept. 2019.
[5] M. Obschonka and D. B. Audretsch, "Artificial intelligence and big data in entrepreneurship: a new era has begun," *Small Business Economics*, pp. 1–11, June 2019.
[6] P. P. Shinde and S. Shah, "A review of machine learning and deep learning applications," in *Proc. Fourth int. conf. comput. commun. control automat. (ICCUBEA), Pune, India*. IEEE, Aug. 2018, pp. 1–6.
[7] P. Li *et al.*, "Multi-key privacy-preserving deep learning in cloud computing," *Future Generation Comput. Syst.*, vol. 74, pp. 76–85, Sept. 2017.
[8] B. McMahan *et al.*, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Int. Conf. Artif. Intell. Statist.* PMLR, Apr. 2017, pp. 1273–1282.
[9] Q. Yang *et al.*, "Federated machine learning: Concept and applications," *ACM Trans. Intell. Syst. and Technol. (TIST)*, vol. 10, no. 2, Jan. 2019.
[10] J. Park, S. Samarakoon, M. Bennis, and M. Debbah, "Wireless network intelligence at the edge," *Proc. IEEE*, vol. 107, no. 11, pp. 2204–2239, Oct. 2019.
[11] P. Kairouz *et al.*, "Advances and open problems in federated learning," *arXiv:1912.04977*, Dec. 2019. [Online]. Available: http://arxiv.org/abs/1912.04977
[12] L. Lyu, H. Yu, and Q. Yang, "Threats to federated learning: A survey," *arXiv:2003.02133*, Mar. 2020. [Online]. Available: https://arxiv.org/abs/2003.02133
[13] T. Li *et al.*, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Process.*, vol. 37, no. 3, pp. 50–60, May 2020.
[14] Z. Du *et al.*, "Federated learning for vehicular internet of things: Recent advances and open issues," *IEEE Open J. Comput. Soc.*, vol. 1, pp. 45–61, May 2020.
[15] M. Aledhari *et al.*, "Federated learning: A survey on enabling technologies, protocols, and applications," *IEEE Access*, vol. 8, pp. 140 699–140 725, July 2020.
[16] V. Kulkarni, M. Kulkarni, and A. Pant, "Survey of personalization techniques for federated learning," in *Proc. Fourth World Conf. Smart Trends Syst., Sec. and Sustain. (WorldS4)*, July 2020.
[17] W. Yang *et al.*, "Federated learning in mobile edge networks: A comprehensive survey," *IEEE Commun. Surv. Tuts.*, vol. 22, no. 3, pp. 2031–2063, July-Sept. 2020.
[18] M. Chen *et al.*, "Wireless communications for collaborative federated learning," *IEEE Communi. Mag.*, vol. 58, no. 12, pp. 48–54, Dec. 2020.
[19] Q. Li *et al.*, "A survey on federated learning systems: vision, hype and reality for data privacy and protection," *arXiv:1907.09693*, Jan. 2021. [Online]. Available: http://arxiv.org/abs/1907.09693
[20] O. A. Wahab *et al.*, "Federated machine learning: Survey, multi-level classification, desirable criteria and future directions in communication and networking systems," *IEEE Commun. Surv. Tuts.*, pp. 1–1, Feb. 2021.
[21] S. Abdulrahman *et al.*, "A survey on federated learning: The journey from centralized to distributed on-site learning and beyond," *IEEE Internet Things J.*, vol. 8, no. 7, pp. 5476–5497, Apr. 2021.
[22] L. U. Khan *et al.*, "Federated learning for internet of things: Recent advances, taxonomy, and open challenges," *IEEE Commun. Surv. & Tuts.*, June 2021.
[23] Z. Yang *et al.*, "Federated learning for 6G: Applications, challenges, and opportunities," *Eng.*, Dec. 2021.
[24] A. Z. Tan *et al.*, "Towards personalized federated learning," *IEEE Trans. Neural Netw. Learn. Syst.*, Mar. 2022.
[25] B. Ghimire and D. B. Rawat, "Recent advances on federated learning for cybersecurity and cybersecurity for federated learning for internet of things," *IEEE Internet Things J.*, June 2022.
[26] M. Hao *et al.*, "Efficient and privacy-enhanced federated learning for industrial artificial intelligence," *IEEE Trans. on Ind. Informat.*, vol. 16, no. 10, pp. 6532–6542, Oct. 2019.
[27] M. Salehi *et al.*, "Hierarchical federated learning across heterogeneous cellular networks," in *Proc. IEEE Int. Conf. Acoust. Speech and Signal Process. (ICASSP), Barcelona, Spain*, May 2020, pp. 8866–8870.
[28] Y. Luo and J. Wang, "Technical introduction of wireless mesh network," *Monitoring and Control (ANMC) Cooperate: Xi'an Technological University (CHINA) West Virginia University (USA) Huddersfield University of UK (UK)*, p. 73, June 2021.
[29] H. B. McMahan *et al.*, "Federated learning of deep networks using model averaging," *arXiv:1602.05629*, Feb. 2016. [Online]. Available: http://arxiv.org/abs/1602.05629
[30] K. Bonawitz *et al.*, "Towards federated learning at scale: System design," *arXiv:1902.01046*, Feb. 2019. [Online]. Available: https://arxiv.org/abs/1902.01046
[31] F. Lai *et al.*, "Oort: Informed participant selection for scalable federated learning," *arXiv preprint arXiv:2010.06081*, Oct. 2020. [Online]. Available: https://arxiv.org/abs/2010.06081
[32] M. R. Sprague *et al.*, "Asynchronous federated learning for geospatial applications," in *Proc. Conf. Mach. Learn. Knowl. Discov. Databases*. Springer, Nov. 2018, pp. 21–28.

[33] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, Sept. 2016. [Online]. Available: https://arxiv.org/abs/1609.04747

[34] T. Li *et al.*, "Federated optimization in heterogeneous networks," *arXiv preprint arXiv:1812.06127*, 2018. [Online]. Available: https://arxiv.org/abs/1812.06127

[35] R. Pathak and M. J. Wainwright, "FedSplit: An algorithmic framework for fast federated optimization," *arXiv preprint arXiv:2005.05238*, May 2020. [Online]. Available: https://arxiv.org/abs/2005.05238

[36] S. Reddi *et al.*, "Adaptive federated optimization," *arXiv preprint arXiv:2003.00295*, Feb. 2020. [Online]. Available: https://arxiv.org/abs/2003.00295

[37] D. Basu *et al.*, "Qsparse-local-SGD: Distributed SGD with quantization, sparsification, and local computations," *arXiv preprint arXiv:1906.02367*, June 2019. [Online]. Available: https://arxiv.org/abs/1906.02367

[38] J. Hamer, M. Mohri, and A. T. Suresh, "Fedboost: A communication-efficient algorithm for federated learning," in *Proc. Int. Conf. Mach. Learn.* PMLR, Nov. 2020, pp. 3973–3983.

[39] M. Al-Quraan *et al.*, "Fedtrees: A novel computation-communication efficient federated learning framework investigated in smart grids," *arXiv preprint arXiv:2210.00060*, Oct. 2022. [Online]. Available: http://arxiv.org/abs/2210.00060

[40] D. Lhuissier, "Etsi - multi-access edge computing - standards for MEC," 2021. [Online]. Available: https://www.etsi.org/technologies/multi-access-edge-computing?jjj=1622661275132

[41] C. S. Wright, "Bitcoin: A peer-to-peer electronic cash system," *Available at SSRN 3440802*, Oct. 2008.

[42] K. Toyoda and A. N. Zhang, "Mechanism design for an incentive-aware blockchain-enabled federated learning platform," in *Proc. IEEE Int. Conf. Big Data (Big Data), Los Angeles, CA, USA*, Dec. 2019, pp. 395–403.

[43] D. C. Nguyen *et al.*, "Federated learning meets blockchain in edge computing: Opportunities and challenges," *IEEE Internet Things J.*, Apr. 2021.

[44] Y. Zhao *et al.*, "Privacy-preserving blockchain-based federated learning for IoT devices," *IEEE Internet Things J.*, vol. 8, no. 3, pp. 1817–1829, Aug. 2020.

[45] M. Aloqaily, I. Al Ridhawi, and M. Guizani, "Energy-aware blockchain and federated learning-supported vehicular networks," *IEEE Trans. Intell. Transp. Syst.*, Aug. 2021.

[46] Y. Lu *et al.*, "Blockchain and federated learning for 5G beyond," *IEEE Netw.*, vol. 35, no. 1, pp. 219–225, Jan./Feb. 2021.

[47] X. Foukas, G. Patounas, A. Elmokashfi, and M. K. Marina, "Network slicing in 5G: Survey and challenges," *IEEE Commun. Mag.*, vol. 55, no. 5, pp. 94–100, May 2017.

[48] B. Brik and A. Ksentini, "On predicting service-oriented network slices performances in 5G: A federated learning approach," in *Proc. 45th IEEE Conf. Local Comput. Netw. (LCN), Sydney, NSW, Australia*, Nov. 2020, pp. 164–171.

[49] H. P. Phyu, D. Naboulsi, and R. Stanica, "Mobile traffic forecasting for network slices: A federated-learning approach," in *Proc. 33rd Int. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*. IEEE, Sept. 2022.

[50] S. Messaoud *et al.*, "Deep federated Q-learning-based network slicing for industrial IoT," *IEEE Trans. Ind. Inform.*, vol. 17, no. 8, pp. 5572–5582, Oct. 2020.

[51] Y. Liu *et al.*, "Device association for RAN slicing based on hybrid federated deep reinforcement learning," *IEEE Trans. Veh. Technol*, vol. 69, no. 12, pp. 15 731–15 745, Dec. 2020.

[52] A. Aissioui *et al.*, "On enabling 5G automotive systems using follow me edge-cloud concept," *IEEE Trans. Veh. Technol.*, vol. 67, no. 6, pp. 5302–5316, Feb. 2018.

[53] T. Subramanya and R. Riggio, "Centralized and federated learning for predictive VNF autoscaling in multi-domain 5G networks and beyond," *IEEE Trans. Netw. Service Manag.*, vol. 18, no. 1, pp. 63–78, Mar. 2021.

[54] Y. Wei *et al.*, "Federated learning empowered end-edge-cloud cooperation for 5G HetNet security," *IEEE Netw.*, vol. 35, no. 2, pp. 88–94, Mar./Apr. 2021.

[55] S. Jere *et al.*, "Federated learning in mobile edge computing: An edge-learning perspective for beyond 5G," *arXiv:2007.08030*, July 2020. [Online]. Available: https://arxiv.org/abs/2007.08030

[56] A. M. Elbir and S. Coleri, "Federated learning for channel estimation in conventional and RIS-assisted massive MIMO," *IEEE Trans. Wireless Commun.*, Nov. 2021.

[57] T. Vu *et al.*, "Energy-efficient massive MIMO for serving multiple federated learning groups," in *Proc. Global Commun. Conf. (GLOBECOM), Madrid, Spain*. IEEE, Dec. 2021, pp. 1–6.

[58] D. Yu *et al.*, "Optimizing over-the-air computation in IRS-aided C-RAN systems," in *21st Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC), Atlanta, GA, USA*. IEEE, May 2020, pp. 1–5.

[59] S. Manzoor *et al.*, "Federated learning empowered mobility-aware proactive content offloading framework for fog radio access networks," *Future Gener. Comput. Syst.*, vol. 133, pp. 307–319, Aug. 2022.

[60] Z. Ji and Z. Qin, "Federated learning for distributed energy-efficient resource allocation," *arXiv preprint arXiv:2204.09602*, Apr. 2022. [Online]. Available: http://arxiv.org/abs/2204.09602

[61] A. Hammoud *et al.*, "AI, blockchain, and vehicular edge computing for smart and secure IoV: Challenges and directions," *IEEE Internet Things Mag.*, vol. 3, no. 2, pp. 68–73, June 2020.

[62] S. R. Pokhrel and J. Choi, "Federated learning with blockchain for autonomous vehicles: Analysis and design challenges," *IEEE Trans. Commun.*, vol. 68, no. 8, pp. 4734–4746, Aug. 2020.

[63] S. Samarakoon, M. Bennis, W. Saad, and M. Debbah, "Distributed federated learning for ultra-reliable low-latency vehicular communications," *IEEE Trans. Commun.*, vol. 68, no. 2, pp. 1146–1159, Nov. 2019.

[64] D. Ye *et al.*, "Federated learning in vehicular edge computing: A selective model aggregation approach," *IEEE Access*, vol. 8, pp. 23 920–23 935, Jan. 2020.

[65] Y. M. Saputra *et al.*, "Federated learning meets contract theory: Economic-efficiency framework for electric vehicle networks," *IEEE Trans. Mobile Comput.*, pp. 1–1, Dec. 2020.

[66] B. Qolomany *et al.*, "Particle swarm optimized federated learning for industrial IoT and smart city services," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2020, pp. 1–6.

[67] A. Rovira-Sugranes *et al.*, "A review of AI-enabled routing protocols for UAV networks: Trends, challenges, and future outlook," *arXiv preprint arXiv:2104.01283*, Apr. 2021. [Online]. Available: http://arxiv.org/abs/2104.01283

[68] Y. Wang *et al.*, "Learning in the air: Secure federated learning for UAV-assisted crowdsensing," *IEEE Trans. Netw. Sci. Eng.*, pp. 1–1, Aug. 2021.

[69] Y. Liu *et al.*, "Federated learning in the sky: Aerial-ground air quality sensing framework with UAV swarms," *IEEE Internet Things J.*, Sept. 2020.

[70] S. Tang *et al.*, "Battery-constrained federated edge learning in UAV-enabled IoT for B5G/6G networks," *arXiv:2101.12472*, Jan. 2021. [Online]. Available: https://arxiv.org/abs/2101.12472

[71] H. Zhang and L. Hanzo, "Federated learning assisted multi-UAV networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 11, pp. 14 104–14 109, Nov. 2020.

[72] W. Ni *et al.*, "Optimal transmission control and learning-based trajectory design for UAV-assisted detection and communication," in *Proc. IEEE 31st Ann. Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC), London, UK*, Oct. 2020, pp. 1–6.

[73] H. A. B. Salameh *et al.*, "Jamming-aware simultaneous multi-channel decisions for opportunistic access in delay-critical IoT-based sensor networks," *IEEE Sensors J.*, vol. 22, no. 3, pp. 2889–2898, Dec. 2021.

[74] N. I. Mowla *et al.*, "Federated learning-based cognitive detection of jamming attack in flying Ad-Hoc network," *IEEE Access*, vol. 8, pp. 4338–4350, Dec. 2019.

[75] E. Basar *et al.*, "Wireless communications through reconfigurable intelligent surfaces," *IEEE Access*, vol. 7, pp. 116 753–116 773, Aug. 2019.

[76] K. Yang *et al.*, "Federated machine learning for intelligent IoT via reconfigurable intelligent surface," *IEEE Netw.*, vol. 34, no. 5, pp. 16–22, Sept. 2020.

[77] H. Liu, X. Yuan, and Y.-J. A. Zhang, "Reconfigurable intelligent surface enabled federated learning: A unified communication-learning design approach," *IEEE Trans. Wireless Commun.*, June 2021.

[78] H. Liu, X. Yuan, and Y. A. Zhang, "CSIT-Free federated edge learning via reconfigurable intelligent surface," *arXiv:2102.10749*, Feb. 2021. [Online]. Available: https://arxiv.org/abs/2102.10749

[79] L. Li *et al.*, "Enhanced reconfigurable intelligent surface assisted mm-Wave communication: A federated learning approach," *Chin. Commun.*, vol. 17, no. 10, pp. 115–128, Oct. 2020.

[80] Y. Shaikh, V. Parvati, and S. Biradar, "Survey of smart healthcare systems using internet of things IoT," in *Proc. IEEE Int. Conf. Commun. Comput. Internet Things (IC3IoT), Chennai, India*, Feb. 2018, pp. 508–513.

[81] L. Romeo *et al.*, "Internet of robotic things in smart domains: Applications and challenges," *Sensors*, vol. 20, no. 12, p. 3355, Jan. 2020.

[82] I. Mohammed *et al.*, "Budgeted online selection of candidate IoT clients to participate in federated learning," *IEEE Internet Things J.*, vol. 8, no. 7, pp. 5938 – 5952, Apr. 2020.

[83] Y. Liu *et al.*, "Deep anomaly detection for time-series data in industrial IoT: A communication-efficient on-device federated learning approach," *IEEE Internet Things J.*, vol. 8, no. 8, pp. 6348 – 6358, Apr. 2020.

[84] Y. Chen *et al.*, "Fedhealth: A federated transfer learning framework for wearable healthcare," *IEEE Intell. Syst.*, vol. 35, no. 4, pp. 83–93, Apr. 2020.

[85] B. Liu *et al.*, "Experiments of federated learning for COVID-19 chest X-ray images," *arXiv:2007.05592*, July 2020. [Online]. Available: https://arxiv.org/abs/2007.05592

[86] W. Yang *et al.*, "FFD: A federated learning based method for credit card fraud detection," in *Proc. Big Data 8th Int. Congr., Services Conf. Federation, (SCF), San Diego, CA, USA*, K. Chen, S. Seshadri, and L. Zhang, Eds., vol. 11514, June 2019, pp. 18–32.

[87] J. Singh *et al.*, "Micro-LED as a promising candidate for high-speed visible light communication," *Appl. Sci.*, vol. 10, no. 20, p. 7384, Jan. 2020.

[88] S. Idris *et al.*, "Visible light communication: A potential 5G and beyond communication technology," in *Proc. 15th Int. Conf. Electron., Comput. Comput. (ICECCO), Abuja, Nigeria*, Dec. 2019, pp. 1–6.

[89] J. Zhang *et al.*, "Cell-free massive MIMO: A new next-generation paradigm," *IEEE Access*, vol. 7, pp. 99 878–99 888, July 2019.

[90] T. L. Marzetta, *Fundamentals of massive MIMO*. Cambridge University Press, Nov. 2016.

[91] T. Tozer and D. Grace, "High-altitude platforms for wireless communications," *Electron. & Commun. Eng. J.*, vol. 13, no. 3, pp. 127–137, June 2001.

[92] L. Lei *et al.*, "Beam illumination pattern design in satellite networks: Learning and optimization for efficient beam hopping," *IEEE Access*, vol. 8, pp. 136 655–136 667, July 2020.

[93] M. Azmat and S. Kummer, "Potential applications of unmanned ground and aerial vehicles to mitigate challenges of transport and logistics-related critical success factors in the humanitarian supply chain," *Asian J. Sustain. Social Responsib.*, vol. 5, no. 1, pp. 1–22, Dec. 2020.

[94] Q. Lan *et al.*, "What is semantic communication? a view on conveying meaning in the era of machine intelligence," *J. Commun. Inform. Netw.*, vol. 6, no. 4, pp. 336–371, Dec. 2021.

[95] V. Balasubramanian *et al.*, "Intelligent resource management at the edge for ubiquitous IoT: an SDN-based federated learning approach," *IEEE netw.*, vol. 35, no. 5, pp. 114–121, Nov. 2021.

[96] A. G. Roy *et al.*, "BrainTorrent: A peer-to-peer environment for decentralized federated learning," *arXiv:1905.06731*, May 2019. [Online]. Available: http://arxiv.org/abs/1905.06731

[97] T. Nishio and R. Yonetani, "Client selection for federated learning with heterogeneous resources in mobile edge," *IEEE Int. Conf. Commun. (ICC)*, pp. 1–7, May 2019.

[98] Y. Zhan *et al.*, "A learning-based incentive mechanism for federated learning," *IEEE Internet Things J.*, vol. 7, no. 7, pp. 6360–6368, Jan. 2020.

[99] J. Kang and Xothers, "Incentive mechanism for reliable federated learning: A joint optimization approach to combining reputation and contract theory," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 10 700–10 714, Sept. 2019.

[100] H. Yang, M. Fang, and J. Liu, "Achieving linear speedup with partial worker participation in non-IID federated learning," *arXiv preprint arXiv:2101.11203*, Jan. 2021. [Online]. Available: http://arxiv.org/abs/2101.11203

[101] Q. Wu, K. He, and X. Chen, "Personalized federated learning for intelligent IoT applications: A cloud-edge based framework," *IEEE Open J. Comput. Soc.*, vol. 1, pp. 35–44, Feb. 2020.

[102] S. Augenstein *et al.*, "Generative models for effective ML on private, decentralized datasets," *arXiv preprint arXiv:1911.06679*, Nov. 2019. [Online]. Available: http://arxiv.org/abs/1911.06679

[103] A. Albaseer *et al.*, "Exploiting unlabeled data in smart cities using federated edge learning," in *Proc. 16th Int. Wireless Commun. Mobile Comput. Conf. (IWCMC), Limassol, Cyprus*, June 2020, pp. 1666–1671.

[104] K. Yang *et al.*, "Federated learning via over-the-air computation," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 2022–2035, Jan. 2020.

[105] G. Zhu, Y. Wang, and K. Huang, "Broadband analog aggregation for low-latency federated edge learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 491–506, Oct. 2019.

[106] G. Zhu *et al.*, "One-bit over-the-air aggregation for communication-efficient federated edge learning: Design and convergence analysis," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 2120–2135, Nov. 2020.

[107] X. Yao, C. Huang, and L. Sun, "Two-stream federated learning: Reduce the communication costs," in *Proc. IEEE Vis. Commun. Imag. Process. (VCIP), Taichung, Taiwan*, Dec. 2018, pp. 1–4.

[108] Y. Lin *et al.*, "Deep gradient compression: Reducing the communication bandwidth for distributed training," *arXiv:1712.01887*, Dec. 2017. [Online]. Available: https://arxiv.org/abs/1712.01887

[109] H. Sun, X. Ma, and R. Q. Hu, "Adaptive federated learning with gradient compression in uplink NOMA," *IEEE Trans. Veh. Technol.*, vol. 69, no. 12, pp. 16 325–16 329, Sept. 2020.

[110] M. Chen *et al.*, "Convergence time optimization for federated learning over wireless networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 4, pp. 2457–2471, Dec. 2020.

[111] T. Huang *et al.*, "Stochastic client selection for federated learning with volatile clients," *arXiv preprint arXiv:2011.08756*, Nov. 2020. [Online]. Available: https://arxiv.org/abs/2011.08756

[112] W. Xia, W. Wen, K.-K. Wong, T. Q. Quek, J. Zhang, and H. Zhu, "Federated-learning-based client scheduling for low-latency wireless communications," *IEEE Wireless Commun.*, vol. 28, no. 2, pp. 32–38, May 2021.

[113] L. N. Smith, "Cyclical learning rates for training neural networks," in *Proc. IEEE winter conf. appl. comput. vision (WACV), Santa Rosa, CA, USA*, Mar. 2017, pp. 464–472.

[114] C. Zhao *et al.*, "Secure multi-party computation: theory, practice and applications," *Inform. Sci.*, vol. 476, pp. 357–372, Feb. 2019.

[115] H. B. McMahan *et al.*, "Learning differentially private recurrent language models," *arXiv preprint arXiv:1710.06963*, Oct. 2017. [Online]. Available: http://arxiv.org/abs/1710.06963

[116] Y. Lu *et al.*, "Differentially private asynchronous federated learning for mobile edge computing in urban informatics," *IEEE Trans. Ind. Informat.*, vol. 16, no. 3, pp. 2134–2143, Mar. 2020.

[117] C. Zhang *et al.*, "BatchCrypt: Efficient homomorphic encryption for cross-silo federated learning," in *Proc. USENIX Annu. Techn. Conf. (ATC)*, July 2020, pp. 493–506.

[118] Y. Sun *et al.*, "Update or wait: How to keep your data fresh," *IEEE Trans. Inform. Theory*, vol. 63, no. 11, pp. 7492–7508, Aug. 2017.

[119] M. Costa, M. Codreanu, and A. Ephremides, "On the age of information in status update systems with packet management," *IEEE Trans. Inform. Theory*, vol. 62, no. 4, pp. 1897–1910, Feb. 2016.

[120] S. Bayhan, G. Gür, and A. Zubow, "The future is unlicensed: Coexistence in the unlicensed spectrum for 5G," *arXiv preprint arXiv:1801.04964*, Jan. 2018. [Online]. Available: http://arxiv.org/abs/1801.04964

[121] T. Hospedales *et al.*, "Meta-learning in neural networks: A survey," *arXiv preprint arXiv:2004.05439*, Apr. 2020. [Online]. Available: http://arxiv.org/abs/2004.05439

[122] A. Jaegle *et al.*, "Perceiver: General perception with iterative attention," *arXiv preprint arXiv:2103.03206*, Mar. 2021. [Online]. Available: http://arxiv.org/abs/2103.03206

[123] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. neural inform. process. syst.*, 2017, pp. 5998–6008.

[124] C.-J. Wu *et al.*, "Sustainable AI: Environmental implications, challenges and opportunities," *arXiv preprint arXiv:2111.00364*, Oct. 2021. [Online]. Available: http://arxiv.org/abs/2111.00364

[125] B. Guler and A. Yener, "Sustainable federated learning," *arXiv preprint arXiv:2102.11274*, Feb. 2021. [Online]. Available: http://arxiv.org/abs/2102.11274

[126] Y. He, X. Zhang, and J. Sun, "Channel pruning for accelerating very deep neural networks," in *Proc. IEEE int. conf. comput. vis. (ICCV)*, 2017, pp. 1389–1397.

[127] S. Chen, W. Wang, and S. J. Pan, "Metaquant: Learning to quantize by learning to penetrate non-differentiable quantization," *Adv. Neural Inform. Process. Syst.*, vol. 32, pp. 3916–3926, 2019.

[128] H. Qin *et al.*, "Binary neural networks: A survey," *Pattern Recognit.*, vol. 105, p. 107281, Sept. 2020.

[129] L. U. Khan *et al.*, "Digital-Twin-Enabled 6G: Vision, architectural trends, and future directions," *arXiv preprint arXiv:2102.12169*, Feb. 2021. [Online]. Available: http://arxiv.org/abs/2102.12169

[130] H. Sami *et al.*, "AI-based resource provisioning of IoE services in 6G: A deep reinforcement learning approach," *IEEE Trans. Net. Service Manage.*, vol. 18, no. 3, pp. 3527–3540, Mar. 2021.

[131] L. L. Pipino, Y. W. Lee, and R. Y. Wang, "Data quality assessment," *Commun. ACM*, vol. 45, no. 4, pp. 211–218, Apr. 2002.