POLITECNICO DI TORINO
Repository ISTITUZIONALE

Handling Privacy-Sensitive Medical Data With Federated Learning: Challenges and Future Directions

Original
Handling Privacy-Sensitive Medical Data With Federated Learning: Challenges and Future Directions / Aouedi, Ons; Sacco, Alessio; Piamrat, Kandaraj; Marchetto, Guido. - In: IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS. - ISSN 2168-2194. - ELETTRONICO. - 27:2(2023), pp. 790-803. [10.1109/JBHI.2022.3185673]

Availability:
This version is available at: 11583/2968639 since: 2023-07-08T09:57:36Z

Publisher:
IEEE

Published
DOI:10.1109/JBHI.2022.3185673

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

29 March 2024

Handling Privacy-Sensitive Medical Data With Federated Learning: Challenges and Future Directions

Ons Aouedi, *Student Member, IEEE*, Alessio Sacco, *Student Member, IEEE*, Kandaraj Piamrat, *Member, IEEE*, and Guido Marchetto, *Senior Member, IEEE*

***Abstract*—Recent medical applications are largely dominated by the application of Machine Learning (ML) models to assist expert decisions, leading to disruptive innovations in radiology, pathology, genomics, and hence modern healthcare systems in general. Despite the profitable usage of AI-based algorithms, these data-driven methods are facing issues such as the scarcity and privacy of user data, as well as the difficulty of institutions exchanging medical information. With insufficient data, ML is prevented from reaching its full potential, which is only possible if the database consists of the full spectrum of possible anatomies, pathologies, and input data types. To solve these issues, Federated Learning (FL) appeared as a valuable approach in the medical field, allowing patient data to stay where it is generated. Since an FL setting allows many clients to collaboratively train a model while keeping training data decentralized, it can protect privacy-sensitive medical data. However, FL is still unable to deliver all its promises and meets the more stringent requirements (e.g., latency, security) of a healthcare system based on multiple Internet of Medical Things (IoMT). For example, although no data are shared among the participants by definition in FL systems, some security risks are still present and can be considered as vulnerabilities from multiple aspects. This paper sheds light upon the emerging deployment of FL, provides a broad overview of current approaches and existing challenges, and outlines several directions of future work that are relevant to solving existing problems in federated healthcare, with a particular focus on security and privacy issues.**

***Index Terms*—Federated learning, Internet of Medical Things, healthcare, privacy.**

I. Introduction

WHILE edge computing, leveraging the multitude of Internet of Things (IoT) devices, is fast booming due to its proficiency in various aspects (e.g., reducing the latency and congestion of network), this emerging paradigm is also guiding new research fields. Among them, we can name smarter healthcare infrastructures, which are favored by recent advances in the Internet of Medical Things (IoMT), including wireless sensors, medical devices, and mobile healthcare [1], [2]. IoMT devices are particularly efficient in monitoring blood pressure, glucose levels, heart rate, and body temperature, as well as providing remote monitoring of patients in a real-time manner. Behind these devices, a large amount of data is collected to be analyzed for a real-time decision such as patient situation [3], [4].

For these tasks, machine learning (ML) and deep learning (DL) models can act in the background to improve the end-user's (e.g., patients) comfort and reduce the possible risks. Concretely, ML and especially DL-based models have led to disruptive innovations in the healthcare-based systems such as chronic disease monitoring [3], cancer prediction [5], tumor detection [6]. Modern DL models feature millions of parameters that need to be learned from sufficiently large curated data sets in order to achieve clinical-grade accuracy while being safe, fair, and generalizing well to unseen data. However, operators face difficulty in obtaining a large amount of data, which would become possible if hospitals were willing to share their sensitive data. The need to protect electronic health records is accentuated by several international regulatory policies, set to restrict data access and protect medical data privacy. For example, the Health Insurance Portability and Accountability Act (HIPAA)[1] in the USA and the General Data Protection Regulation (GDPR) in European Union[2] completely redefine the data management policy. Consequently, valuable data are often confined to individual hospitals and cannot be leveraged for analysis, hindering the application of DL in the healthcare context. Moreover, the increase in heterogeneity of data sources could decrease the performance of the model, bottleneck the whole network, and cause an extra computational cost for both storage and processing.

To leverage the value of existing health datasets while protecting privacy-sensitive patients' data, Federated Learning (FL) appeared as a promising solution. With FL, a global model is trained collaboratively by each agent of the system (e.g., hospitals, IoMTs, or health care centers) over the decentralized

Manuscript received 3 March 2022; revised 3 June 2022; accepted 16 June 2022. Date of publication 23 June 2022; date of current version 6 February 2023. *(Corresponding author: Alessio Sacco.)*

Ons Aouedi and Kandaraj Piamrat are with LS2N, Nantes University 44322, France (e-mail: ons.aouedi@ls2n.fr; kandaraj.piamrat@ls2n.fr).

Alessio Sacco and Guido Marchetto are with DAUIN, Politecnico di Torino, 10129 Turin, Italy (e-mail: alessio_sacco@polito.it; guido.marchetto@polito.it).

Digital Object Identifier 10.1109/JBHI.2022.3185673

[1][Online]. Available: https://www.cdc.gov/phlp/publications/topic/hipaa.html

[2][Online]. Available: https://gdpr-info.eu/issues/data-protection-officer/

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

network. This is done via local updates, without exchanging private data. FL replaces the sharing of medical data across medical institutions, which is most of the time prohibited. With the sharing of local trained model information (e.g., parameters and gradients), it is possible to obtain sufficient knowledge for model training. However, although FL implicitly offers a certain degree of privacy since sensitive data is not exposed and traffic may be encrypted [7], additional effort is needed to ensure that the algorithm is proceeding optimally without compromising security or patient privacy. For example, even if data is anonymized, gathering just a few data attributes may allow patient re-identification [8].

Moreover, FL aims to improve not only privacy but also training efficiency as it uses the computation power and data of potentially millions of IoMT devices/hospitals for training in parallel. Extracting common knowledge from IoMT devices helps to achieve a high-quality global model that guides the automation process. Nevertheless, the increase in heterogeneity of data sources could decrease the performance of the model, bottleneck the whole network, and cause an extra cost for both storage and processing. In this paper, we aim to provide crucial information about the use of FL in healthcare while highlighting unsolved technical questions.

A. Related Work

Driven by the importance of healthcare systems and IoMT, several related reviews have been conducted. For example, Vishnu *et al.* [1] present a brief overview of IoMT-based remote monitoring systems, smart hospitals, mobile health, and how they can be used to improve treatments for chronic diseases. Also, Kagita *et al.* [11] present a short paper on the privacy and security concerns with IoMT systems without discussing possible FL-based solutions.

Recently, the use of DL/ML-based models for healthcare applications has received special attention because of their unique features to improve the quality of services and solve complex problems. In this context, Qayyum *et al.* [9] propose a comprehensive survey on different security challenges related to the application of these data-driven algorithms in healthcare-based systems. Similarly, Qadri *et al.* [13] present a survey on the use of ML for Healthcare-IoT and other relevant technologies, including edge computing, blockchain, Big Data, and software-defined networks.

However, it appears that these solutions based on ML/DL models are obstructed by the scarcity and privacy of user data as well as the difficulty of collecting the data in a central entity. This made FL a good alternative to enhance data confidentiality. As a result, several review papers present the application of FL in many existing domains, including healthcare. For example, Liu *et al.* [22] introduce the application of FL to 6 G networks, while [23] presents an FL approach to route packets in virtualized networks. Moreover, Brik *et al.* [24] present the integration of FL and Unmanned Aerial Vehicles (UAVs)-enabled wireless networks. Most importantly, Xu *et al.* [10] present a concise review on the application of FL with Healthcare. Designed explicitly for IoMT scenarios, [25] presents a cutting-edge FL system that, based on blockchain technology, optimizes the consensus phase by dividing large clusters of IoMT into multiple smaller clusters. Gadekallu *et al.* [20] present a survey on the use of FL for Big Data services and applications.

As for security aspects, Aledhari *et al.* [14] provide a review of the FL architecture and framework, specific review papers focus on the security, threats, and privacy issues related to FL. For instance, Mothukuri *et al.* [12] present a comprehensive survey on the security and privacy concerns with FL-based solutions whereas Lyu *et al.* [26] study the threats to FL, focusing specifically on the poisoning and inference attacks. In the same direction, Agrawal *et al.* [16] provide a review of FL-based approaches for intrusion detection systems along with their challenges and vulnerabilities. Similarly, Ferrag *et al.* [17] present a comprehensive survey, as well as an experimental analysis of FL approaches for cyber security in the Internet of Things (IoT) applications. Furthermore, Ghimire *et al.* [18] present a detailed study on FL and its application in cybersecurity and cybersecurity for FL. Li *et al.* [19] propose a comprehensive survey on the integration of Blockchain and FL. Last but not least, Ali *et al.* [21] presented an overview of the integration of FL and blockchain for IoT applications. Table I summarizes the contributions and limitations of the existing survey.

B. Contribution

Although FL has been well studied in different domains, to the best of our knowledge, no existing work extensively reviews the use of FL in IoMT networks and applications. This motivates us to investigate the integration of FL into healthcare systems and to specifically consider and review the application of FL in IoMT as well as the security and privacy concerns. In brief, the key contributions and novelties of this paper can be summarized as follows:

- **Preliminary discussion of FL in IoMT:** We present the FL concept and discuss the motivations behind the application of FL in IoMT, which include privacy and security issues, latency, communication overhead, and scalability of the healthcare system.
- **FL for IoMT applications:** We review the recent studies on FL with IoMT such as Federated transfer learning, FL with cloud-edge computing, as well as FL for COVID-19 identification.
- **FL for security and privacy concerns:** We describe how FL preserves the privacy and security of the patient. We also review the techniques combined with FL in order to further improve the security and privacy of the IoMT system, such as blockchain and encryption methods.
- **Challenges and future directions on FL for IoMT:** We present issues and challenges suffered from FL in practice, including single point of failure drawback, communication bottleneck, straggler clients, model convergence with non-independent and identically distributed (IID) data, and limited capacity of IoMT devices.

TABLE I
SUMMARY OF RELATED REVIEWS ON FEDERATED LEARNING AND DATA-DRIVEN METHODS FOR HEALTHCARE AND INTERNET OF MEDICAL THINGS

Ref.	Contributions	Federated Environment	IoMT	Security & Privacy in FL
[1]	An overview of IoMT based remote monitoring systems, smart hospitals, mobile health, and IoMT based enhanced chronic disease treatment methods.	x	✓	x
[9]	A comprehensive survey on several security challenges related to the application of ML/DL in healthcare-based systems.	x	✓	x
[10]	An overview on the current progress on FL to healthcare services.	✓	✓	x
[11]	An overview on the privacy and security concerns with IoMT services.	x	✓	x
[12]	An overview on the privacy/security issues and the different defensive proposed solutions for FL.	x	✓	x
[13]	A survey on the future technologies for Healthcare-IoT including ML, edge computing, blockchain, big data, and SDN.	x	✓	x
[14]	An overview on the existing FL architectures and frameworks.	x	*	x
[15]	An extensive review on the FL application in the Industrial Internet of things (IIoT).	✓	*	x
[16]	An extensive review on the use of FL in intrusion detection system.	✓	x	x
[17]	A comprehensive survey and experimental analysis of FL approaches for cyber security in the Internet of Things (IoT) applications.	x	*	✓
[18]	A detailed a detailed study on FL and its application in cybersecurity and cybersecurity for FL.	✓	x	✓
[19]	A comprehensive survey on the integration of blockchain and FL.	✓	*	✓
[20]	A comprehensive survey on the use of FL for big data services and applications.	✓	*	x
[21]	An overview about the integration of FL and blockchain for IoT applications.	✓	x	✓

✓, x, and * indicates that the topic is totally, not or partially covered respectively.

C. Paper Organization

The rest of the paper is organized as follows. In Section II, we briefly describe the workflow of typical FL algorithms, while Section III provides a comprehensive portfolio of current solutions applying FL-based approaches in healthcare and IoMT. Given the importance of privacy and security of patient data, we overview current guarantees and limitations of FL privacy preservation in Section IV. Then, Section V discusses open challenges and issues in using FL for IoMT, showing future research directions. Finally, Section VI concludes our paper.

II. FEDERATED LEARNING CONCEPTS

The need for a huge amount of data for ML/DL model training in healthcare has spawned many initiatives aimed at bringing together data from different institutions. However, medical and patient data may have significant business value, making it less likely that they will be freely shared among medical organizations or with cloud providers. Given the privacy concerns and data governance challenges, FL tackles these issues by enabling collaborative learning without centralizing data. It provides a highly trained ML model without the risk of exposing training data since they are kept where they are generated. While the increasing amount of data has led to the striking success of ML models, data management has also opened new problems that FL can potentially solve. For example, along with the data privacy and delocalization problems, FL also solves the problem of having inadequate data by providing a trust factor between heterogeneous domains. At the same time, this federated environment comes with another benefit of having a model trained on larger landscape data.

As depicted in Fig. 1, FL enables gaining insights collaboratively, e.g., in the form of a consensus model, without moving patient data beyond the firewalls of the institutions in which they reside. Unlike a traditional ML process, the learning phase occurs locally at each participating institution, and only model characteristics (e.g., parameters, weights, gradients) are transferred. While training without FL requires centralized training, in which data acquiring sites donate their data to a central data manager from which they and others are able to extract data for local and independent training, FL is an iterative process, wherein each communication round the model performance can be improved. With FL, training can occur in two different versions: *centralized* and *decentralized*. The former, which corresponds to the typical FL workflow, is characterized by the presence of an aggregation server. In this scenario, a federation of training nodes works locally on the available data; each node submits its partially trained model to a central server intermittently for aggregation and then, using the received global model, continues training on the consensus model returned by the server. Conversely, in a decentralized or peer-to-peer FL, the model aggregation does not require a central entity. Each training node exchanges its partially trained models with some or all of its peers, and each does its own aggregation. Independently of the FL training strategy, a recent analysis for the medical field has shown that a general FL-trained model can achieve performance levels comparable to the ones trained on centrally hosted datasets and superior to models that only see isolated single-institutional data [27], [28].

As can be seen, the FL scenario consists of two main phases: *local update* and *global aggregation*. While the local update relies on common ML/DL models according to the designed method (e.g., DNN, DRL, CNN), the aggregation phase is at the heart of the FL process and aims to reduce the variance of the weight updates affecting the prediction accuracy of the model. Several aggregation mechanisms have been proposed for FL and have shown to achieve acceptable performance: (i) Federated-Averaging *(FedAvg)* [29], the most widely used

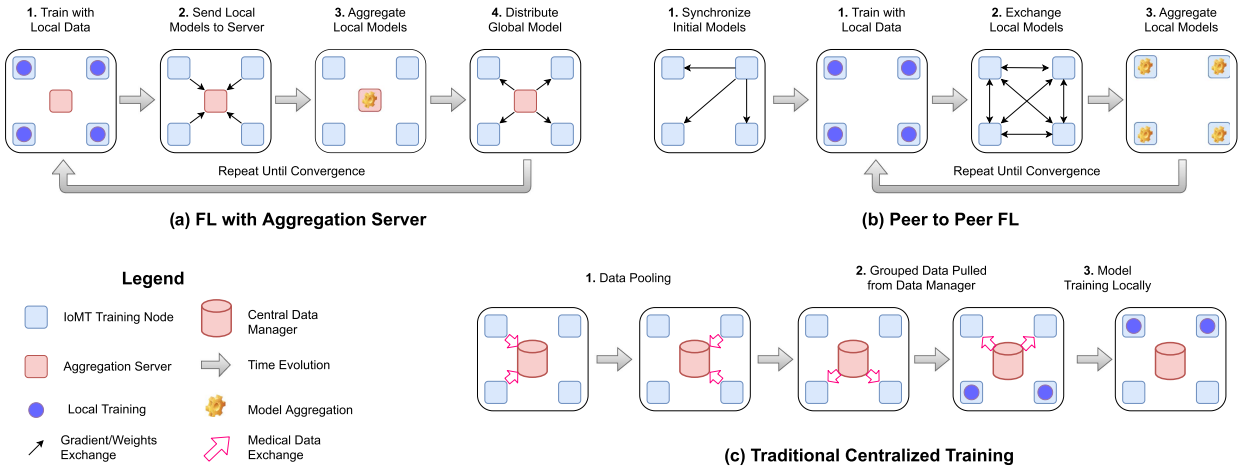


Fig. 1. Typical FL workflows in comparison to a traditional learning based on a centralized data manager. (a) Centralized and (b) peer to peer FL formulations allow private data to remain local to clients. (c) A general non-FL training workflow where data collection agents send their data to a central entity.

algorithm, given its simplicity yet efficacy and robustness, (ii) Adaptive Federated Averaging [30], designed to detect failures, attacks, and bad updates provided by participants, (iii) Part-Data-Sharing strategy [31], which addresses the statistical challenge of FL when local data is non-independent and identically distributed (IID) (see Section V-D) by creating a small subset of data, which is globally shared among all the edge devices, (iv) *FedProx* [32], which extends FL to heterogeneous network conditions by adding a proximal term to the objective that helps improve the stability of the method, and (v) *Qffedavg* [33], that, inspired by fair resource allocation in wireless networks, encourages a fairer (i.e., more uniform) distribution of the model performance across devices in a federated network.

Moreover, data distribution has a significant impact on the FL deployment and the associated practical and technical challenges. In particular, there exist three types of federated learning: *horizontal federated learning*, in which the data sets share the same feature space but differ in the sampling space, *vertical federated learning*, in which the data sets differ in the feature space but share the same sampling space, and *federated transfer learning*, in which the data sets has different feature space as well as different sampling space.

III. FL for Internet of Medical Things

The privacy-preserving feature appears vital for medical data, which are extremely sensitive to the patients and hospitals. Since FL was introduced in order to keep the data where they are generated (IoMT devices) and to preserve the data privacy [7], several solutions have attempted to integrate FL into medical IoT applications.

In the context of electronic health records, for example, FL helps to represent and find clinically similar patients [34], as well as predict hospitalizations due to cardiac events [35]. A novel FL-based clinical decision support system can be found in [36], in which the authors have integrated FL, Recurrent Neural Networks (RNN)-based models, and attention mechanisms in order to provide accurate solutions. The goal of this system is to assist healthcare professionals in medical diagnosing and overcome privacy concerns for sharing sensitive data. In another work, Sozinov *et al.* [37] attempted to use FL for the activity recognition tasks. Then, they study the trade-off between the communication cost and the model accuracy, which indicates that FL requires less communication and computational resources while using less complex models but at the cost of lower accuracy. Also, the authors showed that when the training data is IID, the difference between FL and non-FL is within 3%. In addition, the authors proposed an FL algorithm that identifies and rejects erroneous clients while achieving an accuracy close to FL without erroneous clients. Furthermore, Han *et al.* [38] proposed a zero-watermarking scheme based on FL in order to solve the privacy and security issues of the teledermatology healthcare framework.

Recently, researchers have used FL with traditional ML-based models, due to the complexity of some DL models as well as the limited computation resources of most IoMT devices. For example, Brisimi *et al.* [35] focused on the hospitalization prediction for patients with heart-related diseases using electronic health records. To do so, the authors train a soft-margin support vector machine (SVM) in a collaborative way by keeping every participant's data private. The theoretical and experimental comparisons show that the proposed model converges faster and with less communication overhead compared to an alternative distributed algorithm.

In fact, as labeling data is often difficult and time-consuming, researchers have started to reformulate the FL as a semi-supervised model by combining both supervised learning (using labeled data) and unsupervised learning (no label data) [39]. In this context, Zhao *et al.* [40] proposed semi-supervised FL for human activity recognition where the clients locally train an unsupervised model using their unlabeled data and the server integrates the resulting global unsupervised model into the pipeline of the supervised learning process. Their experimental results show that human activity recognition with semi-supervised FL

is not affected by the non-IID data and can achieve comparable accuracy to that of the supervised FL.

Another area of applicability for FL is within health industrial and collaborative research for companies that even compete with each other. One of the biggest initiatives in this context is the Melloddy project, which aims to apply multi-task FL to the datasets of 10 pharmaceutical companies [41]. By training a predictive model that allows conclusions to be drawn about how chemical compounds bind to proteins, the partners aim to optimize the drug discovery process without revealing their extremely valuable internal data.

Although the efficiency of FL application for healthcare systems, the shared global model trained on the FL server fails in personalization due to the different characteristics or daily activity patterns of users; hence, it is important to have a fine-grained or personalized model. One way to achieve this is via Transfer Learning.

A. Federated Transfer Learning

Healthcare applications are often different but related to each others, making knowledge transfer inter-domains possible and leading to a more precise and personalized model. As a result, the recent technique of transfer learning, which avoids learning from scratch and solves the problem of insufficient training data, can be effective in this scenario [42]. For instance, Chan *et al.* [43] proposed a Federated Transfer Learning (FTL) framework for remote healthcare monitoring, called *FedHealth*, which is the first FTL framework for Human Activity Recognition. At first, the model was aggregated using FL, and then the transfer learning was applied to create personalized models for each organization. Also, this framework uses a homomorphic encryption algorithm to enable secure model sharing between the organizations and the cloud (i.e., the FL server). To evaluate the performance of their solution, the authors used a public human activity recognition dataset called *UCI Smartphone*. The experiment results demonstrate that *FedHealth* can improve the classification performance compared to the non-federated model and the traditional machine learning model. In a similar work, Elayan *et al.* [44] proposed a TFL framework using IoMT devices in order to detect skin diseases. The results demonstrate that the FTL outperforms the non-TFL approach in terms of the Area Under The Curve (AUC) metric and maintains the same accuracy. It also shows that the TFL framework increases the classification time. In addition, to tackle the heterogeneity in IoT environments, Wu *et al.* [45] proposed a personalized FL using transfer learning, called *PerFit*. The results show that the accuracy of PerFit is 11.12% higher than that of classical FL.

B. FL With Cloud-Edge Architecture

Since the IoMT devices (e.g., smartwatches) are limited hardware in terms of storage and computational capabilities, edge computing has been proposed and gained popularity. It is an efficient solution to address these issues and reduce the network congestion and latency that occur with the cloud. As shown in Fig. 2, edge computing helps to train the model closer to the end-users and hence may conduct inference much faster than in the cloud [46]. Note that the edge equipment in the healthcare system can be computers or powerful devices installed in hospitals for model training. In such a context, Hakak *et al.* [47] proposed a general edge-based FL framework. This framework consists of three modules: the cloud module, the edge module, and the application module. The cloud module is managed by the model owner and is used as the FL server, whereas the edge module collects the data from the application module and updates the local model. On the other hand, the application is responsible for the sensor activity. It notifies the edge module when some activity is detected. Moreover, Wu *et al.* [46] proposed an FL-based framework for in-home monitoring of health using a cloud-edge architecture, called *FedHome*. Any user of an IoMT device in *FedHome* can offload the local model training task to the home edge. Then, the cloud collects the edge local model parameters using the *FedAvg* algorithm. This framework takes advantage of the feature extraction ability of convolutional neural network (CNN), dimension reduction capability of autoencoder, and oversampling strategy of synthetic minority over-sampling technique (SMOTE) to cope with the imbalanced, non-IID, and communication overhead issues. The experimental results on human activity recognition demonstrate the effectiveness of the *FedHome* framework in terms of accuracy, computation, and communication overhead.

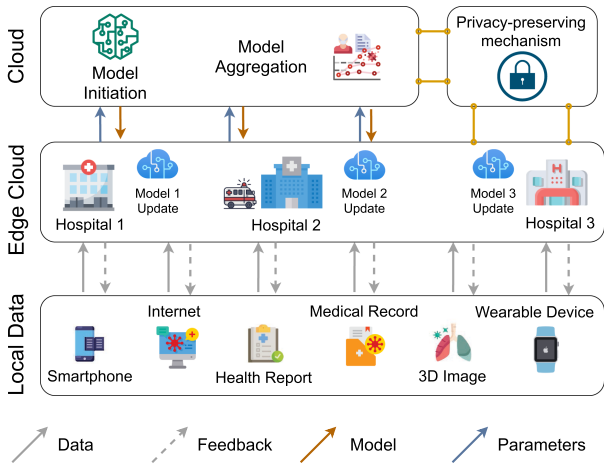


Fig. 2. FL framework in edge computing system for healthcare services. In this hierarchical and centralized setting, the edge cloud aggregates data used for training the model, whose logic resides in the central cloud.

C. FL for COVID-19

The COVID-19 pandemic has caused an unprecedented global crisis [48]. During the period of the COVID-19 pandemic, FL has recently been used to contain the virus spread, given its ability to detect the positive case by training the models of isolated medical institutions. For example, Yan *et al.* [49] focused on COVID-19 chest X-ray images using different non-federated/federated learning-based models including MobileNet, ResNet-18, ResNeXt and COVID-Net. The experimental results show that ResNet-18 has the best performance

in training both with FL and without FL approaches, while ResNeXt shows the highest efficiency on images with COVID-19 labels. As a result, the work recommends using ResNet-18 and ResNeXt models for COVID-19 identification. In a similar work, Feki *et al.* [50] used FL to classify X-ray images into COVID-19 infected cases and non-COVID-19 ones. To do so, they adopted two well-known CNN architectures, namely VGG-16 and ResNet-50, in a collaborative way. Then, FL-VGG-16 and FL-ResNet-50 were studied under different settings such as the training data size, the number of clients participating in each round, IID/Non-IID, and balanced/unbalanced data. The experimental results demonstrate that FL can achieve comparable performance to the methods without FL under these settings without sacrificing the privacy of the end-users. In the same direction, Wang *et al.* [51] proposed a 5G-enabled architecture for COVID-19 diagnosis. Specifically, multiple hospitals collaborate through FL with user privacy preservation. Qayyum *et al.* [52] proposed a clustered FL framework in order to tackle the convergence issues of the classical FL model due to the diverse distribution of the data. Specifically, the proposed framework trains a multi-modal ML model in a collaborative way using both X-ray and Ultrasound imagery. Furthermore, since the non-IID and the imbalanced data can decrease the performance of FL, Nguyen *et al.* [53] proposed *FedGAN* in order to address dataset limitation and imbalance data issues. *FedGAN* is a novel scheme for COVID-19 detection obtained by enabling the joint design of FL and a Generative Adversarial Network (GAN) in a federated way. Specifically, *FedGAN* aims to achieve better COVID-19 image augmentation where each client trains the generator and discriminator in order to compute the local gradients.

In summary, we list FL-IoMT approaches in Table II to recapitulate the ML/DL models, the used datasets as well as the key contributions and limitations of each approach.

D. Summary and Discussion

From the presented state-of-the-art on FL-IoMT, we can notice that FL plays an important role in facilitating healthcare services. It also improves patient privacy and reduces low latency during the collaboration of multiple hospitals/entities [7]. It can be learned from [37] that FL requires less communication overhead, and hence it can replace traditional learning approaches in IoMT applications. Moreover, recently FL has made a huge contribution to the fight against COVID-19 without sharing patient data. Also, based on [43], transferring the knowledge in distributed healthcare system provides a high-quality personalized model. FL also requires computationally robust devices. For this reason, some IoMT devices offload their model training task to the edge gateways; therefore, helping the FL training process to take advantage of the cloud and edge computing, and using the IoMT devices as simple data collectors. Furthermore, as shown in Table II, the majority of the works have considered the CNN model and its variations, such as ResNet and VGG. Since finding a suitable model is not an easy task, a comparative analysis to preliminary evaluate the performance of other ML models is one of the best practices for developing future FL architecture.

IV. Security and Privacy Concerns

Since federated learning comes with a privacy-preserving attribute, it can play a significant role in various industry domains that involve sensitive personal data, such as IoMT for healthcare. Medical data is highly sensitive and must be protected by means of appropriate confidentiality procedures. For example, in a collaborative healthcare scenario, each hospital or medical research center holds sensitive diagnostic data that cannot be shared with others; however, they desire to learn from each other's data. Although FL is the enabling technology in this scenario and provides a privacy-preservation capability by allowing the clients to keep the data on local devices, there are still model security and data leakage risks that would compromise the security of the FL system and the data privacy of clients. In the following subsections, some important issues are raised and discussed.

A. How Insecure is Federated Learning?

Before describing various attacks leading to privacy leakage in FL systems, we briefly summarize the characteristics of security and privacy problems in the following.

Security problem (data and model manipulation): This type of problem is primarily caused by curious or malicious attackers targeting vulnerabilities of the FL system, which can lead to significant performance drop and sometimes model invalidation. Clearly, this process is extremely hazardous and could negatively affect thousands of devices. In the context of healthcare and IoMT scenarios, an attacker can directly manipulate the model and data of a local affiliation, resulting in a malicious update of the global model or mislabeled data.

Privacy problem: This type of problem, even more severe than the security one, arises when vulnerabilities cause user data leakage, as it weakens the basics of FL that are designed explicitly for privacy preservation across multi-device ML. For instance, if messages carrying the global model and local gradient updates exchanged between the central server and a local device are intercepted, gradient-based reconstruction attack algorithms can be applied to recover the raw data in the local device. In IoMT applications, the intercepted data could be a patient's personal or healthcare information, which presents a severe ethical problem. In particular, private data can be extracted indirectly even from the shared information, by means of some emerging techniques, such as model inversion of the model updates [56], gradients themselves [57] and adversarial attacks [58]. FL, despite a different training process compared to traditional ML, still suffers from information leakage issues and, having multiple parties, extends the attack surface. An adversary can observe changes over time, or specific model updates, i.e., updates of a single agent, to reverse engineer and obtain some knowledge about data. For instance, Carlini *et al.* [59] demonstrate that it can be extracted sensitive text patterns, e.g., a specific credit card number, from a recurrent neural network trained on users' language data. Besides, the model can be manipulated by attackers inducing additional memorization through gradient-ascent-style attacks.

TABLE II
OVERVIEW OF RECENT STUDIES ON FEDERATED LEARNING WITH IOMT

Ref.	Algorithm	Datasets	Contribution	Limitations
[35]	SVM	Boston Medical Center	The authors train a soft-margin support vector machine (SVM) in a collaborative way by keeping every participant's data private	The non-IID condition was not explored.
[36]	RNN	Disease-Symptom Knowledge Database	The authors exploited the FL and attention mechanism to train an RNN-based model collaboratively by learning a shared global model at the FL server.	The model's complexity was not discussed.
[37]	DNN, Softmax	The Heterogeneity Human Activity Recognition Dataset	The authors find that FL could achieve comparable accuracy to centralized learning. Also, they demonstrate that FL requires less communication and compute resources when using less complex models, but at the cost of lower accuracy.	Small scale experiment.
[40]	Autoencoder, LSTM	Opp, DG, PAMAP2	The authors demonstrate that semi-supervised FL is not affected by the non-IID data and can achieve comparable accuracy to that supervised FL.	The performance under the heterogeneity of devices constrains was not explored.
[43]	CNN	UCI Smartphone	The average accuracy of the FedHealth framework is 5.3% and 4% better than non-federated learning and no transfer approaches, respectively.	The model's performance under the IoMT devices' computing heterogeneity was not addressed.
[44]	ResNet-50	Atlas Dermatology dataset	The FTL framework outperforms the non-TFL approach in terms of the AUC (Area Under The Curve) metric and maintaining the same accuracy.	The heterogeneity of IoMT devices was not studied.
[45]	CNN	MobiAct dataset	The authors show that transfer FL can outperform the classical FL in terms of accuracy as well as the centralized learning in terms of communication overhead.	The model complexity was not discussed.
[46]	Convolutional Autoencoder	MobiAct dataset	FedHome framework takes advantage of the feature extraction ability of CNN, dimension reduction capability of autoencoder, and oversampling strategy of SMOTE to cope with the imbalanced, non-IID, and communication overhead issues.	Only labeled data have been used during the training process.
[47]	N/A	N/A	The authors proposed an FL-based framework for leveraging Edge computing to support healthcare analytics based on user-generated data.	The proposed work was not validated on a real dataset.
[49]	MobileNet, ResNet-18, MoblieNet, COVID-Net	Covid-chestxray-dataset	A comparative analysis of different deep learning-based models with and without a federated framework.	The non-IID condition was not explored.
[50]	VGG-16, ResNet-50	Covid-chestxray-dataset, Public chest X-ray dataset	Study the performance of the FL-based methods under several factors: the training data size, the number of clients participating in each round, IID/Non-IID, and balanced/unbalanced data.	Small scale experiment (i.e., only four clients).
[51]	DNN	N/A	The authors proposed a A 5G-enabled architecture for COVID-19 diagnosis using FL mechanism.	The non-IID condition was not explored.
[52]	VGG-16	COVID-19 CT segmentation dataset, chest ultrasound images	The authors proposed a clustered FL framework in order to tackle the convergence issues of the classical FL model due to the diverse distribution of the data.	The non-IID condition was not explored.
[53]	CNN,GAN	DarkCOVID, ChestCOVID	FedGAN trains a GAN in a collaborative way to generate realistic COVID-19 images to improve the COVID-19 detection.	The performance in a heterogeneity environment was not tested.
[54]	GhostNet, ResNet-50, ResNet-101	Covid-chestxray-dataset, COVID-CT	The proposed system consists of two important points which are client participation and client selection in order to reduce the training time, communication cost while maintaining the detection performance.	The non-IID constraint was not explored.
[55]	DNN	PhysioNet 2017	A novel FL designed by transferring activation and gradients for the forward and backward propagation instead of the model parameters in order to decrease the communication overhead.	The non-IID constraint was not explored.
[38]	Sparse Autoencoder	N/A	The authors proposed a zero-watermarking scheme based on FL in order to solve the privacy and security issues of the teledermatology healthcare framework.	The complexity of the model was not discussed.

Although some countermeasures exist, e.g., limiting the granularity of updates and adding noise [27], and ensuring adequate differential privacy [60], the study of effective methods to improve the security and privacy protection is still an active area of research [61]. At the same time, it can be noted how some sophisticated countermeasures can be avoided if all parties are deemed trustworthy. For FL consortia in which all parties are bounded by an enforceable collaboration agreement, we can disregard some of the more nefarious motivations, such as deliberated attempts to extract sensitive information or to intentionally corrupt the model. This would lead to the mere principles of standard collaborative research. However, operating FL systems on a larger scale might be impractical to establish collaborative agreements and it is more reasonable not to have trustworthiness assumptions. Some clients, for example, may deliberately try to extract information from other parties, degrade performance, or bring the system down. Security strategies that mitigate these risks can include advanced encryption of model submissions, secure authentication of all parties, actions traceability, differential privacy, verification systems, execution integrity, model confidentiality, and protection against adversarial attacks.

Beyond providing rigorous privacy guarantees, it is necessary to develop methods that are computationally resource-effective, communication efficient, and tolerant to dropped devices, without overly compromising accuracy. Despite the variety of privacy definitions in FL, typically, they generally fall into two

categories: global privacy and local privacy. The former requires that the model updates generated at each round are private to all untrusted third parties except the central server, while the latter further requires that the updates are also private to the server. In other words, in a global privacy-enhancing mechanism, the global server is assumed trusted, while in local privacy, the central server may be malicious.

Lastly, privacy problems are exacerbated in medical companies by strict laws and regulations designed to prevent the risk of re-identification and data breaches. In fact, personal health information is considered to be highly sensitive, containing not only diagnostic and healthcare related information but also identifiable details about individuals. Even from the consumer perspective, data privacy is one of the public's most important concerns because data breaches can result in reduced public trust. For these reasons, any FL-based data management system must comply with these laws.

While standards are necessary for security and privacy purposes, they can make sharing and using health data challenging. A trivial example of possible applications that can be limited includes sharing of information between medical centers, hospitals, and governments. In view of the above arguments, precious data often remain confined to individual institutions and are rarely leveraged for analysis, hindering the application of deep learning algorithms in healthcare. If these standards are respected, FL can foster the application of deep learning in digital health and healthcare informatics, making it possible the use of existing health datasets while protecting user privacy.

B. Existing Countermeasures and Solutions

Built upon previous classical *cryptographic protocols*, such as secure multiparty computation, Bonawitz *et al.* [62] introduce a secure aggregation protocol for protecting individual model updates. In this solution, the central server is not able to see any local updates, but it can still observe the exact aggregated results at each round. The applied secure aggregation is a lossless method that can retain the original accuracy with a very high privacy guarantee. The main drawback of the resulting method, however, is the significant extra communication cost. Other approaches are able to offer global privacy by applying *differential privacy* to federated learning, as in [63]. This type of approach comes with a considerable number of hyperparameters that need to be carefully selected and that affect communication and accuracy. For stronger privacy guarantees, Bhowmick *et al.* [64] have proposed a relaxed version of local privacy by limiting the power of potential adversaries. Providing stronger guarantees than global privacy, it has better model performance than strict local privacy. Li *et al.* [65] introduced locally differentially private algorithms in the context of meta-learning while providing provable learning guarantees in convex settings. Meta-learning techniques are based on the sharing of knowledge gained from individual learning tasks to catalyze the learning of similar unseen tasks and can be applied to federated learning with personalization. Moreover, differential privacy can be combined with model-compression algorithms to reduce communication and simultaneously obtain privacy benefits [66].

An initial attempt to enhance the innovation and creative capability of health-related organizations, while guaranteeing the fulfillment of specific medical laws, has been proposed in [67]. The article presents an open *innovation framework* in the healthcare industry, namely Open Health, by building a next-generation collaborative framework with partner organizations and the research community. Specifically designed for healthcare data, Hao *et al.* [68] have proposed a privacy-aware and resource-saving collaborative learning protocol, called *PRCL*. In this work, the authors have partitioned the model, i.e., neural network, into three parts: the first and the last parts are trained on the client-side, while the middle part, which is the heavy one, is outsourced to be trained on the cloud servers. Before training the model on the client-side, the training data are perturbed by adding Gaussian noise, and packets are secured via homomorphic encryption to efficiently perform gradients aggregation in the ciphertext context. The simulation results show that PRCL reduces the local training overhead because outsourcing the middle part to the cloud server while providing similar accuracy to other state-of-the-art approaches. At the same time, confidentiality is preserved since the cloud cannot access the plaintext gradient and an attacker can only get the perturbed data due to the added noise. Similar results have been achieved by *EPPDA* [69], an efficient privacy-preserving *data aggregation method* for FL that, based on secret sharing to resist the reverse attack, can covertly aggregate user-trained models without disclosing the user model.Adopting the homomorphisms of secret sharing, it preserves user privacy and requires less computing and communication resources than alternatives as [62], and is able to provide acceptable fault tolerance in the event of user disconnection. To protect the model and data manipulation, Wang *et al.*proposed a defensive strategy based on comparing the uploaded model's accuracy and size [70]. If lower the respective target values, the lower the trust value of the client, and when such a trust value is lower than 60%, the request from this client will be prohibited.

Another approach to establish data security and eliminate the trust issues is the deployment of FL on *blockchain* technology [71]–[73]. The main advantage brought is the replacement of the central authority with a specially designed decentralized privacy protocol. The blockchain architecture has a specially designed distributed ledger structure that connects blocks in chronological order, allowing to share and maintain saved data in all nodes in a decentralized environment [74]. In addition, the blockchain-based verification scheme can also improve the reliability of the federated training process. Fig. 3 sketches the main components of an FL architecture based on blockchain and highlights the differences from a traditional FL setting. Compared to a vanilla FL, a blockchain FL consists of devices and miners, where the latter are either randomly selected devices or separate nodes such as network edges that are relatively free from energy constraints in the mining process. Once the agent has computed and uploaded the local model update to its associated miner in the blockchain network, miners verify all the local updates and run the Proof-of-Work. When such a Proof-of-Work is completed, the miner generates a block that records the verified local model and this is added to a blockchain, also known as a distributed ledger, then downloaded by devices.

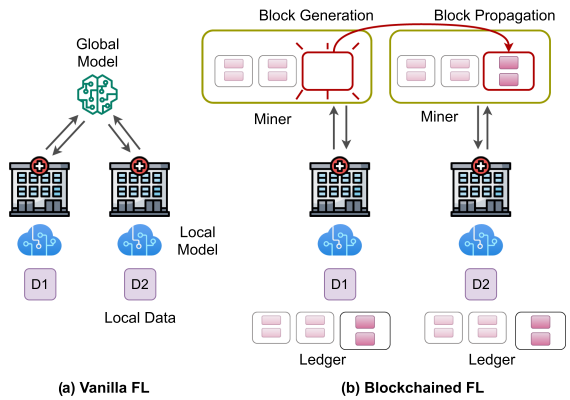


Fig. 3. Blockchained FL architecture *vs.* a vanilla centralized FL system.

Each device is thus able to compute the global model update from the new block. Such a solution offers several advantages such as decentralization, non-tempering, open autonomy, and anonymous traceability [75]. A blockchain-based FL, however, should mitigate the overhead of frequent communications in the blockchain consensus, which can cause excessive latency. For example, [76] presents *TrustFed*, which uses Industrial IoT devices as federated learning candidates and smart contract technology over the Ethereum blockchain to maintain participants' reputations. The solution can also identify and eliminate outliers in the training distributions before combining the model updates. Lu *et al.* [77] designed a new system, called *DITEN*, that integrates blockchain and FL in edge networks and uses Deep Neural Networks (DNN) as a strategy scheduler to ensure the privacy of user data and enhance learning security.

Beside this architecture, other techniques can be considered. For example, to prevent an arbitrary client from trying to reconstruct the private data of another client by exploiting the global model, client-level differential privacy can be achieved even in the context of blockchain. In a general FL model, any single client's update can be hidden by adding random Gaussian noise to the aggregated global model [63]. For blockchained systems, each client could locally add a certain amount of Gaussian noise after local gradient descent steps and submit the model to the blockchain. To also protect the model from the public, the aggregated global model on blockchain could be encrypted with a decryption key held only by the participating clients. To empower FL with this approach, Qu *et al.* [72] propose to store only the pointer of the global updates on-chain, while a distributed hash table is used to save the data. In such a way, block generation efficiency is guaranteed, enabling decentralized privacy protection while preventing single point failure. In a similar way, Lu *et al.* [78] integrated federated learning into the consensus algorithm of the blockchain to save the hash rate in the context of industrial IoT, in which blockchain enables secure data retrieval and ensures accurate model training. Rahman *et al.* [79], with a lightweight hybrid blockchain-based FL framework, propose to perform additive encryption in the edge nodes, while multiplicative encryption in the blockchain aggregates the updated model parameters.

C. Summary and Discussion

In conclusion, the features of FL make it uniquely suited for sensitive data as in the context of IoMT systems, but the architecture and model sharing must address further challenges compared to other fields. For example, when designing an architecture for FL, it is recommended to correctly identify the vulnerabilities of the proposed FL system and prevent unauthorized access by curious or malicious attackers. Implementing such prerequisites to defend against loopholes will help develop a more secure system. Therefore, equipping the solution with the mentioned solutions for backdoor and gradient attacks is a mandatory step for an FL engineer to enhance security and privacy defenses.

Some recent techniques can partially solve these privacy issues [61], however, the more secure the techniques, the lower the accuracy of the final model may be [80]. Thus, it appears crucial to find a trade-off between performance and security guarantees. Other recent methods aiming to enhance the privacy of federated learning use tools such as secure multiparty computation or differential privacy. These approaches often provide privacy at the cost of reduced model performance or system efficiency [62]. Balancing these trade-offs, both theoretically and empirically, is a considerable challenge in realizing private FL systems and especially IoMT-based architectures.

V. Challenges and Solutions

As it can be seen, unremitting research activity on the use of the FL concept for IoMT has produced many novel and interesting solutions. However, in practice, these solutions suffer from several issues and challenges that should be considered when designing FL-based approaches for healthcare services. In this section, we discuss the main challenges that should be considered and the future directions of using the FL concept.

A. Central Server Failure & Robustness via Decentralization

The presence of a centralized server in most existing FL schemes increases the risk of data leakage, especially in distributed multi-parties applications. We can mention two main obstacles: *(i)* a high volume of aggregated data from different parties to be processed by the server; *(ii)* none of these parties fully trust each others (including the server), thus fearing data leakage. As an alternative, FL can operate in decentralized or peer-to-peer topologies, where devices communicate only with their neighbors, as shown in Section II. These decentralized architectures, which remove the need for a central server, can effectively improve the resilience of the training process by eliminating a possible single-point-of-failure. However, while this distributed approach can help to make the system highly scalable, it also opens up some issues in gradient synchronization. For example, some clients may take much longer to report their output than other nodes, and these agents are often referred to as stragglers.

Existing solutions. To mitigate the impact of stragglers in a distributed system, recent work has proposed deadline-based

approaches where all workers compute the local gradients using a variable number of samples within a fixed global cycle [81]. Another line of work proposes asynchronous decentralized SGD, in which the workers update their models based on the last iterates received from their neighbors [82]. Although these asynchronous methods are inherently robust to stragglers, they may suffer from slow convergence due to the use of stale models. Moreover, decentralizing can be combined with blockchain technology, where a blockchain-based approach can jointly solve the problem of single-point-of-failure and central trust. While blockchain removes the need for a trusted curator and connects each participant through multiparty data retrieval, it also strengthens the data-sharing scheme and provides some fault-tolerant guarantees [83].

With either solution, i.e., peer-to-peer or blockchain, the problem is converted to the pursuit of fault tolerance for the clients. Such a property has been extensively studied by the systems community and is a fundamental consideration of classical distributed systems, with some specific studies for ML workloads in data center environments [84]. Given the importance of this problem, specific solutions can be combined with the previous approach, as mentioned in Section V-C.

B. Trade-Off Between Communication Overhead and Accuracy

Despite the fact that data generated on each IoMT remain local and raw data are not sent, the communication is a critical bottleneck in FL networks [85]. Federated networks of IoMT can indeed comprise a massive number of devices (e.g., sensors), and communication in the network can be many orders of magnitude slower than local computation due to limited resources, such as bandwidth, energy, and power [86]. Therefore, a lightweight model is desirable to reduce the communication overhead during the training round as well as developing communication-efficient methods that iteratively send small messages or model updates as part of an affordable FL training process. Two key aspects must be considered for an effective reduction of the communication overhead: the total number of communication rounds, and the size of messages transmitted in each round. Recent solutions attempting to solve this communication bottleneck can be categorized into three classes according to the proposed approach: *(i) local updating methods*, *(ii) compression schemes*, and *(iii) decentralized training*.

Local updating methods are recent approaches proposed to improve communication efficiency in distributed settings by allowing for a variable number of local updates to be applied on each machine in parallel at each communication round, contrary to simply performing mini-batch optimization locally and then aggregating mini-batch updates centrally [87]. The most commonly used method for FL, *FedAvg* [29], is based on averaging local stochastic gradient descent (SGD) updates for the primal problem. While *FedAvg* has been shown to work well for non-convex problems, it does not exhibit convergence guarantees and can diverge in practical settings when data are heterogeneous [32] (see Section V-D for how to address this problem).

Although a high number of local updating methods can effectively reduce the total number of communication rounds, it can be convenient to reduce the size of messages exchanged at each round via *compression schemes* model, e.g., sparsification, subsampling, and quantization. While these methods found applicability in general distributed systems, recent studies have provided practical strategies specific to FL, such as forcing the updating models to be sparse and low rank [88], performing quantization with structured random rotations [88], and using lossy compression and dropout to reduce server-to-device communication [89]. However, convergence guarantees still need to be explored for low device participation and local updating optimization methods.

Lastly, *decentralized training* is a potential alternative in FL to reduce high communication costs on the central server. Although some recent articles have investigated decentralized training over heterogeneous data with local updating schemes, e.g., [90], they are either restricted to linear models or assume full device participation.

Existing solutions. Given the complexity brought by these techniques, recent solutions have expanded these general approaches to further optimize the accuracy of the final model, paving the way for enhanced FL-based healthcare architectures. For instance, Sozinov *et al.* [37] studied and evaluated performance as well as the communication costs in the case of human activity recognition of both softmax regression and DNN for different data distributions. Results confirm that using less complex models, such as softmax regression, is a viable solution for most real-world applications since user behavior phenomena are often simple enough to be captured by relatively simple models. However, for medical data collected by IoMT simple models can be likely inaccurate. Konen *et al.* [88] proposed two types of updating the local client's model before communicating it to a central server: structured and sketched updates. While using a structured update, the client maps the original local model to a lower-dimensional space, via a sketched update, the client compresses it by using, for example, a probabilistic quantization. The authors showed how using CNNs and LSTMs, the communication costs can be reduced by two orders of magnitude compared to the original federated learning algorithm.

Moreover, Jeong *et al.* [91] presented Federated Distillation *(FD)*, a distributed model training algorithm whose communication payload size is much smaller than that of a benchmark FL scheme, particularly when the model size is large. On top of conventional periodic communication of FL, the proposed *FD* exchanges, not the model parameters but the model output, so that large local models can be adopted on the ML device. To further improve the transmission of model output, the approach is combined with Federated Augmentation *(FAug)*, a data augmentation scheme that can collectively learn the trade-off between privacy leakage and communication overhead using a Generative Adversarial Network (GAN).

In conclusion, the most valuable techniques will need to demonstrate improvements at the Pareto frontier, i.e., they must achieve higher accuracy than any other approach under the same communication budget and, possibly, for a wide range of communication/accuracy profiles. It is also still necessary

to compare communication-reduction techniques for FL in a meaningful way, as it has been studied for efficient inference in neural networks [92].

C. Clients Synchronization and Fault Tolerance

When learning over multiple remote devices as in FL, fault tolerance becomes a critical aspect as it is common for some participating devices to drop out at some points before the completion of the given training iteration [85]. IoMT devices from remote areas may be proner to drop due to poor network connectivity or battery failure and, therefore, the trained federated model will be biased towards devices with favorable network conditions. In addition, while a large number of clients can increase the diversity of the local model across them, this may degrade the performance of FL [93]. To preserve effective participation of the clients in the FL process, there are several approaches for fault tolerance and selecting clients when some of them are slow to respond.

Existing solutions. Recent studies present a joint-announcement protocol that can randomly select some devices/clients from a large number of participants in a given training round [85]. Such a joint-announcement protocol is useful to alleviate the out-of-sync issue and tolerate the dropout of some clients at some points before the completion of each training iteration. Moreover, Nishio *et al.* [94] have proposed a new protocol, called *FedCS*. Unlike the classical FL, with *FedCS* the central server is not only used for the model aggregation task but also collects resource information about the client, such as wireless channel states, computational capacities, data size, and amount of observation for each class. Using this information, the FL server decides which client can participate in the training process. By selecting the devices with favorable resources, *FedCS* provides high classification accuracy in a significantly shorter time compared to the classical FL.

In a similar occurrence of device failure, one simple yet practical strategy is to simply ignore such a client, as in the widely used *FedAvg* [85], which may introduce bias in the learning process. However, while several recent studies have investigated convergence guarantees of variants of FL methods [95]–[97], little work has evaluated the impact of failures over the FL algorithm or studied directly the effect of dropped devices. *FedProx* [32], for instance, tolerates that any selected device to perform partial work, in compliance with the underlying system's constraints, and can safely incorporate these partial updates via a proximal term.

Another effective approach to tolerate device failures is named *coded computation*, which implies algorithmic redundancy. Recent work has explored the use of codes to speed up distributed machine learning training, such as [98]–[101]. In the presence of stragglers, gradient coding, and variants [98], [99], [102], the replication of data blocks (or also the gradient computation on those data blocks) across computing nodes leads either to exact or inexact recovery of the true gradients. However, despite being attractive for FL, these methodologies face fundamental challenges in federated networks since sharing data/replication across devices is often infeasible due to privacy constraints and the scale of the network. Therefore, new schemes ensuring fault tolerance would improve the deployment of FL in healthcare where IoMT devices are likely to a fault, encounter challenging network conditions, and exhaust their batteries.

D. Convergence Guarantees for Non-IID Data

Devices often generate and collect data that are not identically distributed across the network, such as in medical devices collecting varied user data related to biological information [53]. In addition to the variety of data, the number of data samples can vary significantly from device to device, and there may be an underlying statistical structure that captures the relationship among devices and their associated distributions. This paradigm of data generation violates the frequently-used assumptions of independent and identically distributed (IID) data and may cause problems in modeling, theoretical analysis, and empirical evaluation of solutions. In [31], the authors experienced a significant reduction in accuracy of up to 50% in the presence of highly skewed non-IID data.

Existing solutions. To solve this issue, a possible solution includes slightly changing the FL problem from the canonical form (i.e., learning a single global model) to the alternative of learning distinct local models simultaneously via *multi-task learning* frameworks [103]. Alongside, most known approaches in federated learning can be combined with *meta-learning*, i.e., machine learning algorithms that learn from the output of other machine learning algorithms [65]. Both approaches (multi-task and meta-learning) enable personalized or device-specific modeling, which is considered a valuable attempt to handle the statistical heterogeneity of the data.

Alternatively, another simple yet effective approach referred to as Part-Data-Sharing has been presented in [31]. It addresses the statistical challenge of FL when local data is non-IID by distributing a small amount of globally shared data containing examples from each class. However, despite introducing a trade-off between accuracy and centralization, this could lead to significant communication overhead, especially for medical data. To rectify the non-IID training dataset while addressing the communication overhead, Jeong *et al.* [91] proposed *FAug*, in which each device can generate the missing data samples locally using a generative model, i.e., a conditional GAN. A central server trains the generator, while each device detects the labels missing in data samples and uploads a few seed data samples of these target labels to the server. The server performs oversampling of the uploaded seed data samples, e.g., via Google image search for visual data, to train the GAN model. Once the trained generator of the GAN is downloaded, each device can replenish the target labels until an IID training dataset is achieved

E. FL With Constrained IoMT Devices

The need to spread healthcare services over larger areas may lead to the use of very small IoMT devices, often referred to as nano-devices, whose size has a few hundred nanometers [104]. Their utilization can make healthcare services more personalized, and these applications can take

place in unprecedented locations in non-invasive ways (e.g., intra-body). Some IoMT devices, even though not limited in size, come with limitations in the available on-board resources. Their limited power and computational resources make the participation of such constrained IoMT devices in the FL process almost impossible. For example, training CNN-based models for COVID-19 detection (e.g., ResNet-50) requires storage and memory resources that may not be available on all devices. A possible solution in which edge computing provides complementary resources in support of the local model training [105], would require time-consuming data transmission between the nano-devices and the edge nodes and could lead to network congestion, especially given the huge amount of data generated by these devices every second. In such a context, efficient and fast local model training on the nano-devices is needed, and interactive healthcare services also demand extremely low latency. To address these issues, the computational operations need to be lowered, for example by utilizing some lightweight ML/DL models that can run onboard nano-devices.

Existing solutions. Caldas *et al.* [89] introduce *Federated Dropout* to reduce the computation load of local training. Specifically, they have been zeroing out some fixed number of activation at each layer and hence get less complex model computation costs. This solution was inspired by the well-known idea of dropout [106]. Simulation results demonstrate the superior performance of this solution by reducing the local computations by $1.7\times$ without affecting the accuracy of the model. Similarly, Xu *et al.* [107] proposed a resource-aware federated learning framework, called *ELFISH*. With *ELFISH*, the neurons that have less significant weight parameter updates will be randomly masked. These neurons are masked in a single cycle and will recover themselves in the subsequent cycles. In the same direction, Jiang *et al.* [108] proposed a new FL paradigm called *PruneFL* in order to minimize the model's size. The experimental results show that PruneFL always converges to similar accuracy achieved by classical FL. Furthermore, Anh *et al.* [109] applied a Deep Q-learning algorithm based on the Double Deep Q-Network (DDQN) to optimize resource allocation for model training. At each iteration, the FL server needs to decide how much data and energy each mobile device uses to train the model to minimize the total energy consumption and the training latency while meeting the requirements of the training tasks. The simulation results show that the proposed approach can minimize energy consumption and improve the training latency. In addition, data pre-processing is an important step to remove the redundancy in the training set as well as to find the relevant features to be used during the model training [110], [111].

VI. Conclusion

While machine learning approaches, and especially deep learning, are becoming the de-facto knowledge discovery approach in digital healthcare, it appears that data-driven medicine aiming to improve patient care globally requires federated efforts. With the promise of powerful, accurate, safe, robust, and unbiased models, federated learning can neatly protect sensitive medical data while collaboratively training a learning model. Already today, FL is improving medical image analysis by providing clinicians with better diagnostic tools, helping find similar patients for true precision medicine, and optimizing the drug discovery process by reducing costs and time-to-market for pharmaceutical companies. However, the current settings of FL have not answered all the technical questions, and this paper outlined the state-of-the-art approaches and the limitations associated with them. Challenges raised in this paper range from the privacy and security issues to the client synchronization and the presence of non-IID datasets. As such, future FL systems for medical data need to consider further optimizations and network connectivity problems to define an efficient gradient synchronization protocol that can run on limited devices, as often are the IoMT.

References

[1] S. Vishnu, S. J. Ramson, and R. Jegan, "Internet of Medical Things (IOMT)-An overview," in *Proc. 5th Int. Conf. Devices, Circuits Syst.*, 2020, pp. 101–104.

[2] A. Sacco, F. Esposito, G. Marchetto, G. Kolar, and K. Schwetye, "On edge computing for remote pathology consultations and computations," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 9, pp. 2523–2534, Sep. 2020.

[3] O. Aouedi, M. A. B. Tobji, and A. Abraham, "An ensemble of deep auto-encoders for healthcare monitoring," in *Proc. Int. Conf. Hybrid Intell. Syst.*, 2018, pp. 96–105.

[4] A. Sacco et al., "LiveMicro: An edge computing system for collaborative telepathology," in *Proc. 2nd USENIX Workshop Hot Top. Edge Comput.*, 2019. [Online]. Available: https://www.usenix.org/conference/hotedge19/presentation/sacco

[5] Y. Xiao, J. Wu, Z. Lin, and X. Zhao, "A semisupervised deep learning method based on stacked sparse auto-encoder for cancer prediction using RNA-seq data," *Comput. Methods Programs Biomed.*, vol. 166, pp. 99–105, 2018.

[6] L. Zhen and A. K. Chan, "An artificial intelligent algorithm for tumor detection in screening mammogram," *IEEE Trans. Med. Imag.*, vol. 20, no. 7, pp. 559–567, Jul. 2001.

[7] N. Rieke et al., "The future of digital health with federated learning," *NPJ Digit. Med.*, vol. 3, no. 1, pp. 1–7, 2020.

[8] L. Rocher, J. M. Hendrickx, and Y.-A. De Montjoye, "Estimating the success of re-identifications in incomplete datasets using generative models," *Nature Commun.*, vol. 10, no. 1, pp. 1–9, 2019.

[9] A. Qayyum, J. Qadir, M. Bilal, and A. Al-Fuqaha, "Secure and robust machine learning for healthcare: A survey," *IEEE Rev. Biomed. Eng.*, vol. 14, pp. 156–180, 2021.

[10] J. Xu, B. S. Glicksberg, C. Su, P. Walker, J. Bian, and F. Wang, "Federated learning for healthcare informatics," *J. Healthcare Informat. Res.*, vol. 5, no. 1, pp. 1–19, 2021.

[11] M. K. Kagita, N. Thilakarathne, T. R. Gadekallu, and P. K. R. Maddikunta, "A review on security and privacy of Internet of Medical Things," *Intell. Internet Things Healthcare Industry*, pp. 171–187, 2022.

[12] V. Mothukuri, R. M. Parizi, S. Pouriyeh, Y. Huang, A. Dehghantanha, and G. Srivastava, "A survey on security and privacy of federated learning," *Future Gener. Comput. Syst.*, vol. 115, pp. 619–640, 2021.

[13] Y. A. Qadri, A. Nauman, Y. B. Zikria, A. V. Vasilakos, and S. W. Kim, "The future of healthcare Internet of Things: A survey of emerging technologies," *IEEE Commun. Surv. Tut.*, vol. 22, no. 2, pp. 1121–1167, Apr.–Jun. 2020.

[14] M. Aledhari, R. Razzak, R. M. Parizi, and F. Saeed, "Federated learning: A survey on enabling technologies, protocols, and applications," *IEEE Access*, vol. 8, pp. 140699–140725, 2020.

[15] P. Boopalan et al., "Fusion of federated learning and industrial Internet of Things: A survey," *Comput. Netw.*, 2022, Art. no. 109048.

[16] S. Agrawal et al., "Federated learning for intrusion detection system: Concepts, challenges and future directions," 2021, *arXiv:2106.09527*.

[17] M. A. Ferrag, O. Friha, L. Maglaras, H. Janicke, and L. Shu, "Federated deep learning for cyber security in the Internet of Things: Concepts, applications, and experimental analysis," *IEEE Access*, vol. 9, pp. 138509–138542, 2021.

[18] 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.*, vol. 9, no. 11, pp. 8229–8249, Jun. 2022.

[19] D. Li et al., "Blockchain for federated learning toward secure distributed machine learning systems: A systemic survey," *Soft Comput.*, vol. 26, no. 9, pp. 4423–4440, 2022.

[20] T. R. Gadekallu, Q.-V. Pham, T. Huynh-The, S. Bhattacharya, P. K. R. Maddikunta, and M. Liyanage, "Federated learning for Big Data: A survey on opportunities, applications, and future directions," 2021, *arXiv:2110.04160*.

[21] M. Ali, H. Karimipour, and M. Tariq, "Integration of blockchain and federated learning for Internet of Things: Recent advances and future challenges," *Comput. Secur.*, vol. 108, 2021, Art. no. 102355.

[22] Y. Liu et al., "Federated learning for 6G communications: Challenges, methods, and future directions," *China Commun.*, vol. 17, no. 9, pp. 105–118, 2020.

[23] A. Sacco, F. Esposito, and G. Marchetto, "A federated learning approach to routing in challenged SDN-enabled edge networks," in *Proc. 6th IEEE Conf. Netw. Softwarization*, 2020, pp. 150–154.

[24] B. Brik, A. Ksentini, and M. Bouaziz, "Federated learning for UAVs-enabled wireless networks: Use cases, challenges, and open problems," *IEEE Access*, vol. 8, pp. 53841–53849, 2020.

[25] H. Jin, X. Dai, J. Xiao, B. Li, H. Li, and Y. Zhang, "Cross-cluster federated learning and blockchain for internet of medical things," *IEEE Internet Things J.*, vol. 8, no. 21, pp. 15776–15784, Nov. 2021.

[26] L. Lyu, H. Yu, and Q. Yang, "Threats to federated learning: A survey," 2020, *arXiv:2003.02133*.

[27] W. Li et al., "Privacy-preserving federated brain tumour segmentation," in *Proc. Int. Workshop Mach. Learn. Med. Imag.*, 2019, pp. 133–141.

[28] M. J. Sheller, G. A. Reina, B. Edwards, J. Martin, and S. Bakas, "Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation," in *Proc. Int. MICCAI Brainlesion Workshop*, 2018, pp. 92–104.

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

[30] L. Muñoz-González, K. T. Co, and E. C. Lupu, "Byzantine-robust federated machine learning through adaptive model averaging," 2019, *arXiv:1909.05125*.

[31] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-IID data," 2018, *arXiv:1806.00582*.

[32] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proc. Mach. Learn. Syst.*, 2020, vol. 2, pp. 429–450.

[33] T. Li, M. Sanjabi, A. Beirami, and V. Smith, "Fair resource allocation in federated learning," 2019, *arXiv:1905.10497*.

[34] J. Lee et al., "Privacy-preserving patient similarity learning in a federated environment: Development and analysis," *JMIR Med. Informat.*, vol. 6, no. 2, 2018, Art. no. e7744.

[35] T. S. Brisimi, R. Chen, T. Mela, A. Olshevsky, I. C. Paschalidis, and W. Shi, "Federated learning of predictive models from federated electronic health records," *Int. J. Med. Informat.*, vol. 112, pp. 59–67, 2018.

[36] C. M. Thwal, K. Thar, Y. L. Tun, and C. S. Hong, "Attention on personalized clinical decision support system: Federated learning approach," in *Proc. IEEE Int. Conf. Big Data Smart Comput.*, 2021, pp. 141–147.

[37] K. Sozinov, V. Vlassov, and S. Girdzijauskas, "Human activity recognition using federated learning," in *Proc. Int. Conf. Parallel Distributed Process. Appl., Ubiquitous Comput. Commun., Big Data Cloud Comput., Social Comput. Netw., Sustainable Comput. Commun.*, 2018, pp. 1103–1111.

[38] B. Han et al., "Application of robust zero-watermarking scheme based on federated learning for securing the healthcare data," *IEEE J. Biomed. Health Informat.*, Oct. 29, 2021, doi: 10.1109/JBHI.2021.3123936.

[39] O. Aouedi, K. Piamrat, G. Muller, and K. Singh, "Federated semi-supervised learning for attack detection in industrial Internet of Things," *IEEE Trans. Ind. Informat.*, Mar. 7, 2022, doi: 10.1109/TII.2022.3156642.

[40] Y. Zhao, H. Liu, H. Li, P. Barnaghi, and H. Haddadi, "Semi-supervised federated learning for activity recognition," 2020, *arXiv:2011.00851*.

[41] E. Cordis, "Machine learning ledger orchestration for drug discovery," 2019. [Online]. Available: https://cordis.europa.eu/project/id/831472

[42] A. Khamparia, D. Gupta, V. H. C. de Albuquerque, A. K. Sangaiah, and R. H. Jhaveri, "Internet of health things-driven deep learning system for detection and classification of cervical cells using transfer learning," *J. Supercomputing*, vol. 76, no. 11, pp. 8590–8608, 2020.

[43] Y. Chen, X. Qin, J. Wang, C. Yu, and W. Gao, "FedHealth: A federated transfer learning framework for wearable healthcare," *IEEE Intell. Syst.*, vol. 35, no. 4, pp. 83–93, Jul./Aug. 2020.

[44] H. Elayan, M. Aloqaily, and M. Guizani, "Deep federated learning for IoT-based decentralized healthcare systems," in *Proc. Int. Wireless Commun. Mobile Comput.*, 2021, pp. 105–109.

[45] 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, 2020.

[46] Q. Wu, X. Chen, Z. Zhou, and J. Zhang, "FedHome: Cloud-edge based personalized federated learning for in-home health monitoring," *IEEE Trans. Mobile Comput.*, Dec. 16, 2020, doi: 10.1109/TMC.2020.3045266.

[47] S. Hakak, S. Ray, W. Z. Khan, and E. Scheme, "A framework for edge-assisted healthcare data analytics using federated learning," in *Proc. Int. Conf. Big Data*, 2020, pp. 3423–3427.

[48] O. A. Sarumi, O. Aouedi, and L. J. Muhammad, "Potential of deep learning algorithms in mitigating the spread of COVID-19," in *Understanding COVID-19: The Role of Computational Intelligence*. Berlin, Germany: Springer, 2022, pp. 225–244.

[49] B. Yan et al., "Experiments of federated learning for COVID-19 chest X-ray images," in *Proc. Int. Conf. Artif. Intell. Secur.*, 2021, pp. 41–53.

[50] I. Feki, S. Ammar, Y. Kessentini, and K. Muhammad, "Federated learning for COVID-19 screening from chest X-ray images," *Appl. Soft Comput.*, vol. 106, 2021, Art. no. 107330.

[51] R. Wang, J. Xu, Y. Ma, M. Talha, M. S. Al-Rakhami, and A. Ghoneim, "Auxiliary diagnosis of COVID-19 based on 5G-enabled federated learning," *IEEE Netw.*, vol. 35, no. 3, pp. 14–20, May/Jun. 2021.

[52] A. Qayyum, K. Ahmad, M. A. Ahsan, A. Al-Fuqaha, and J. Qadir, "Collaborative federated learning for healthcare: Multi-modal COVID-19 diagnosis at the edge," 2021, *arXiv:2101.07511*.

[53] D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, and A. Y. Zomaya, "Federated learning for COVID-19 detection with generative adversarial networks in edge cloud computing," *IEEE Internet Things J.*, vol. 9, no. 12, pp. 10257–10271, Jun. 2022.

[54] W. Zhang et al., "Dynamic fusion-based federated learning for COVID-19 detection," *IEEE Internet Things J.*, vol. 8, no. 21, pp. 15884–15891, Nov. 2021.

[55] B. Yuan, S. Ge, and W. Xing, "A federated learning framework for healthcare iot devices," 2020, *arXiv:2005.05083*.

[56] B. Wu et al., "P3SGD: Patient privacy preserving SGD for regularizing deep CNNs in pathological image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2099–2108.

[57] L. Zhu and S. Han, "Deep leakage from gradients," in *Federated learning*. Berlin, Germany: Springer, 2020, pp. 17–31.

[58] Z. Wang, M. Song, Z. Zhang, Y. Song, Q. Wang, and H. Qi, "Beyond inferring class representatives: User-level privacy leakage from federated learning," in *Proc. IEEE Conf. Comput. Commun.*, 2019, pp. 2512–2520.

[59] N. Carlini, C. Liu, Ú. Erlingsson, J. Kos, and D. Song, "The secret sharer: Evaluating and testing unintended memorization in neural networks," in *Proc. 28th USENIX Secur. Symp. (USENIX Secur. 19). USENIX Assoc.*, 2019, pp. 267–284.

[60] M. Abadi et al., "Deep learning with differential privacy," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2016, pp. 308–318.

[61] P. Kairouz et al., "Advances and open problems in federated learning," *Found. Trends® Mach. Learn.*, vol. 14, no. 1-2, pp. 1–210, 2021.

[62] K. Bonawitz et al., "Practical secure aggregation for privacy-preserving machine learning," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2017, pp. 1175–1191.

[63] H. B. McMahan, D. Ramage, K. Talwar, and L. Zhang, "Learning differentially private recurrent language models," 2017, *arXiv:1710.06963*.

[64] A. Bhowmick, J. Duchi, J. Freudiger, G. Kapoor, and R. Rogers, "Protection against reconstruction and its applications in private federated learning," 2018, *arXiv:1812.00984*.

[65] J. Li, M. Khodak, S. Caldas, and A. Talwalkar, "Differentially private meta-learning," 2019, *arXiv:1909.05830*.

[66] N. Agarwal, A. T. Suresh, F. X. X. Yu, S. Kumar, and B. Mcmahan, "CPSGD: Communication-efficient and differentially-private distributed SGD," *Adv. Neural Inf. Process. Syst.*, vol. 31, pp. 7564–7575, 2018.

[67] G. Long, T. Shen, Y. Tan, L. Gerrard, A. Clarke, and J. Jiang, "Federated learning for privacy-preserving open innovation future on digital health," *Humanity Driven AI*, Springer, pp. 113–133, 2022.

[68] M. Hao, H. Li, G. Xu, Z. Liu, and Z. Chen, "Privacy-aware and resource-saving collaborative learning for healthcare in cloud computing," in *Proc. IEEE Int. Conf. Commun.*, 2020, pp. 1–6.

[69] J. Song, W. Wang, T. R. Gadekallu, J. Cao, and Y. Liu, "EPPDA: An efficient privacy-preserving data aggregation federated learning scheme," *IEEE Trans. Netw. Sci. Eng.*, Feb. 25, 2022, doi: 10.1109/TNSE.2022.3153519.

[70] W. Wang et al., "Secure-enhanced federated learning for ai-empowered electric vehicle energy prediction," *IEEE Consum. Electron. Mag.*, Sep. 30, 2021 doi: 10.1109/MCE.2021.3116917.

[71] R. Kumar et al., "Blockchain-federated-learning and deep learning models for COVID-19 detection using CT imaging," *IEEE Sensors J.*, vol. 21, no. 14, pp. 16301–16314, Jul. 2021.

[72] Y. Qu et al., "Decentralized privacy using blockchain-enabled federated learning in fog computing," *IEEE Internet Things J.*, vol. 7, no. 6, pp. 5171–5183, Jun. 2020.

[73] K. Salah, M. H. U. Rehman, N. Nizamuddin, and A. Al-Fuqaha, "Blockchain for AI: Review and open research challenges," *IEEE Access*, vol. 7, pp. 10127–10149, 2019.

[74] W. Wang, H. Xu, M. Alazab, T. R. Gadekallu, Z. Han, and C. Su, "Blockchain-based reliable and efficient certificateless signature for IIoT devices," *IEEE Trans. Ind. Informat.*, May 28, 2021, doi: 10.1109/TII.2021.3084753.

[75] T. Ahram, A. Sargolzaei, S. Sargolzaei, J. Daniels, and B. Amaba, "Blockchain technology innovations," in *Proc. IEEE Technol. Eng. Manage. Conf.*, 2017, pp. 137–141.

[76] M. H. ur Rehman, A. M. Dirir, K. Salah, E. Damiani, and D. Svetinovic, "TrustFed: A framework for fair and trustworthy cross-device federated learning in IIoT," *IEEE Trans. Ind. Informat.*, vol. 17, no. 12, pp. 8485–8494, Dec. 2021.

[77] Y. Lu, X. Huang, K. Zhang, S. Maharjan, and Y. Zhang, "Communication-efficient federated learning and permissioned blockchain for digital twin edge networks," *IEEE Internet Things J.*, vol. 8, no. 4, pp. 2276–2288, Feb. 2021.

[78] Y. Lu, X. Huang, Y. Dai, S. Maharjan, and Y. Zhang, "Blockchain and federated learning for privacy-preserved data sharing in industrial IoT," *IEEE Trans. Ind. Informat.*, vol. 16, no. 6, pp. 4177–4186, Jun. 2020.

[79] M. A. Rahman, M. S. Hossain, M. S. Islam, N. A. Alrajeh, and G. Muhammad, "Secure and provenance enhanced internet of health things framework: A blockchain managed federated learning approach," *IEEE Access*, vol. 8, pp. 205071–205087, 2020.

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

[81] N. Ferdinand, H. Al-Lawati, S. C. Draper, and M. Nokleby, "Anytime Minibatch: Exploiting stragglers in online distributed optimization," 2020, *arXiv:2006.05752*.

[82] S. Dutta, G. Joshi, S. Ghosh, P. Dube, and P. Nagpurkar, "Slow and stale gradients can win the race: Error-runtime trade-offs in distributed SGD," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2018, pp. 803–812.

[83] H. Kim, J. Park, M. Bennis, and S.-L. Kim, "Blockchained on-device federated learning," *IEEE Commun. Lett.*, vol. 24, no. 6, pp. 1279–1283, Jun. 2020.

[84] A. Qiao, B. Aragam, B. Zhang, and E. Xing, "Fault tolerance in iterative-convergent machine learning," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 5220–5230.

[85] K. Bonawitz et al., "Toward federated learning at scale: System design," in *Proc. Mach. Learn. Syst.*, vol. 1, pp. 374–388, 2019.

[86] C. Van Berkel, "Multi-core for mobile phones," in *Proc. Des., Automat. Test Europe Conf. Exhib.*, 2009, pp. 1260–1265.

[87] S. Zhang, A. E. Choromanska, and Y. LeCun, "Deep learning with elastic averaging SGD," *Adv. Neural Inf. Process. Syst.*, vol. 28, pp. 685–693, 2015.

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

[89] S. Caldas, J. Konečny, H. B. McMahan, and A. Talwalkar, "Expanding the reach of federated learning by reducing client resource requirements," 2018, *arXiv:1812.07210*.

[90] L. He, A. Bian, and M. Jaggi, "Cola: Decentralized linear learning," *Adv. Neural Inf. Process. Syst.*, vol. 31, pp. 4536–4546, 2018.

[91] E. Jeong, S. Oh, H. Kim, J. Park, M. Bennis, and S.-L. Kim, "Communication-efficient on-device machine learning: Federated distillation and augmentation under non-IID private data," 2018, *arXiv:1811.11479*.

[92] T. Bolukbasi, J. Wang, O. Dekel, and V. Saligrama, "Adaptive neural networks for efficient inference," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 527–536.

[93] Z. Zhang, Z. Yao, Y. Yang, Y. Yan, J. E. Gonzalez, and M. W. Mahoney, "Benchmarking semisupervised federated learning,"2020, *arXiv:2008.11364*.

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

[95] P. Jiang and G. Agrawal, "A linear speedup analysis of distributed deep learning with sparse and quantized communication," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 2530–2541.

[96] S. Wang et al., "Adaptive federated learning in resource constrained edge computing systems," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1205–1221, Jun. 2019.

[97] H. Yu, R. Jin, and S. Yang, "On the linear speedup analysis of communication efficient momentum SGD for distributed non-convex optimization," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 7184–7193.

[98] Z. Charles and D. Papailiopoulos, "Gradient coding using the stochastic block model," in *Proc. IEEE Int. Symp. Inf. Theory*, 2018, pp. 1998–2002.

[99] Z. Charles, D. Papailiopoulos, and J. Ellenberg, "Approximate gradient coding via sparse random graphs," 2017, *arXiv:1711.06771*.

[100] K. Lee, M. Lam, R. Pedarsani, D. Papailiopoulos, and K. Ramchandran, "Speeding up distributed machine learning using codes," *IEEE Trans. Inf. Theory*, vol. 64, no. 3, pp. 1514–1529, Mar. 2018.

[101] A. Reisizadeh, S. Prakash, R. Pedarsani, and A. S. Avestimehr, "Coded computation over heterogeneous clusters," *IEEE Trans. Inf. Theory*, vol. 65, no. 7, pp. 4227–4242, Jul. 2019.

[102] R. Tandon, Q. Lei, A. G. Dimakis, and N. Karampatziakis, "Gradient coding: Avoiding stragglers in distributed learning," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 3368–3376.

[103] V. Smith, C.-K. Chiang, M. Sanjabi, and A. S. Talwalkar, "Federated multi-task learning," *Adv. Neural Inf. Process. Syst.*, vol. 30, pp. 4424–4434, 2017.

[104] N. A. Ali and M. Abu-Elkheir, "Internet of nano-things healthcare applications: Requirements, opportunities, and challenges," in *Proc. IEEE 11th Int. Conf. Wireless Mobile Comput., Netw. Commun.*, 2015, pp. 9–14.

[105] A. Das and T. Brunschwiler, "Privacy is what we care about: Experimental investigation of federated learning on edge devices," in *Proc. First Int. Workshop Challenges Artif. Intell. Mach. Learn. Internet Things*, 2019, pp. 39–42.

[106] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[107] Z. Xu, Z. Yang, J. Xiong, J. Yang, and X. Chen, "ELFISH: Resource-aware federated learning on heterogeneous edge devices," *Ratio*, vol. 2, no. r1, 2019.

[108] Y. Jiang et al., "Model pruning enables efficient federated learning on edge devices," *IEEE Trans. Neural Netw. Learn. Syst.*, 2022.

[109] T. T. Anh, N. C. Luong, D. Niyato, D. I. Kim, and L.-C. Wang, "Efficient training management for mobile crowd-machine learning: A deep reinforcement learning approach," *IEEE Wireless Commun. Lett.*, vol. 8, no. 5, pp. 1345–1348, Oct. 2019.

[110] A. Sacco, F. Esposito, and G. Marchetto, "RoPE: An architecture for adaptive data-driven routing prediction at the edge," *IEEE Trans. Netw. Service Manag.*, vol. 17, no. 2, pp. 986–999, Jun. 2020.

[111] M. Hasan, M. M. Islam, M. I. I. Zarif, and M. Hashem, "Attack and anomaly detection in IOT sensors in IoT sites using machine learning approaches," *Internet Things*, vol. 7, 2019, Art. no. 100059.