

# Federated Learning for Big Data: A Survey on Opportunities, Applications, and Future Directions

Thippa Reddy Gadekallu, Quoc-Viet Pham, Thien Huynh-The,  
Sweta Bhattacharya, Praveen Kumar Reddy Maddikunta, and Madhusanka Liyanage

**Abstract**—Big data has remarkably evolved over the last few years to realize an enormous volume of data generated from newly emerging services and applications and a massive number of Internet-of-Things (IoT) devices. The potential of big data can be realized via analytic and learning techniques, in which the data from various sources is transferred to a central cloud for central storage, processing, and training. However, this conventional approach faces critical issues in terms of data privacy as the data may include sensitive data such as personal information, governments, banking accounts. To overcome this challenge, federated learning (FL) appeared to be a promising learning technique. However, a gap exists in the literature that a comprehensive survey on FL for big data services and applications is yet to be conducted. In this article, we present a survey on the use of FL for big data services and applications, aiming to provide general readers with an overview of FL, big data, and the motivations behind the use of FL for big data. In particular, we extensively review the use of FL for key big data services, including big data acquisition, big data storage, big data analytics, and big data privacy preservation. Subsequently, we review the potential of FL for big data applications, such as smart city, smart healthcare, smart transportation, smart grid, and social media. Further, we summarize a number of important projects on FL-big data and discuss key challenges of this interesting topic along with several promising solutions and directions.

**Index Terms**—Big Data, Federated Learning, Smart City, Smart Healthcare, Smart Transportation

## I. INTRODUCTION

Due to the digitization and advancements in several information and communication technologies like Internet of Things (IoT), communication technologies, smart cities, industrial IoT, stock market, etc., an exponential growth is seen in the data traffic globally. It is expected that the big data market may reach around 230 billion \$ by the year 2025. Almost all the verticals of industries like entertainment, retail, media, manufacturing, healthcare, social media, etc. benefit from the big data [1]. Big data has attracted a great deal of attraction from from academia and industry

Thippa Reddy Gadekallu, Sweta Bhattacharya, Praveen Kumar Reddy Maddikunta are with School of Information Technology and Engineering, Vellore Institute of Technology, India (e-mail: {thippareddy.g, sweta.b, praveenkumarreddy}@vit.ac.in).

Quoc-Viet Pham is with Korean Southeast Center for the 4th Industrial Revolution Leader Education, Pusan National University, Busan 46241, Korea (e-mail: vietpq@pusan.ac.kr).

Thien Huynh-The is with ICT Convergence Research Center, Kumoh National Institute of Technology, Republic of Korea. (e-mail: thienht@kumoh.ac.kr).

Madhusanka Liyanage is School of Computer Science, University College Dublin, Ireland and Centre for Wireless Communications, University of Oulu, Finland, Ireland (e-mail: madhusanka@ucd.ie and madhusanka.liyanage@oulu.fi).

in recent years. Big data has a potential to solve many of the challenging problems in several sectors. The big datasets can be analyzed to uncover patterns through data analytics that attracted the researchers. Some of the examples where big data can play a very important role include using the consumer data, radio frequencies, data from social media, and global position systems for improving several sectors in urban life such as energy, transportation, education, etc. in a smart city, predicting the patients with high risk of cardiac arrests and other health related risks in real time through predictive analytics using the large volumes of electronic health records, etc [2]. Even though big data is used prominently in today's world, there is still a vagueness in the definition of "big data". According to European Commission, big data is large quantity of several types of data generated from different sources like machines, people, or sensors. The data generated may be related to videos, posts in social media, digital images, satellite imagery, and climate information [3]. Big data datasets are of tremendously large sizes with huge dimensions, that cannot be captured, stored, managed, or processed by traditional database technologies [4]. Several characteristics of big data include volume, velocity, veracity, variety [5].

The true potential of big data can be realized only when the value is extracted from massive data by data analytics; where machine learning (ML) plays a very important role due to its ability to understand the patterns and provide insights by learning from the data [6]. ML algorithms can be categorized as supervised ML and unsupervised ML. In supervised ML, the algorithms know both the inputs as well as outputs in prior and the ML algorithms learn how to map inputs to the outputs. Classification and regression are examples of supervised ML. In unsupervised ML, the outputs are not known to the ML algorithms but they discover the patterns within the data on their own. Clustering is an example for unsupervised ML [7]. Traditionally, the data gathered from several big data sources is transferred to a central cloud, where the ML algorithms are trained to understand the patterns from the stored data. This approach of training the big data in a central cloud faces several challenges such as exposing of sensitive information (privacy preservation), incurring additional costs in terms of communication and resources, data management and increased latency which make the real-time analytics very difficult [8], [9]. Among these challenges privacy and security are considered to be very important as sensitive data related to people, vehicles, governments, and organizations can be exposed [10].

Federated learning (FL) is a recent development in ML,

where, instead of transferring of the raw data captures from several sources to the central model located at the server and then training the ML algorithms, the global ML algorithm itself is offloaded to the devices [11]. In FL, only the parameters from the local devices are transferred to the central ML algorithm for global training and predictions [12]. This property of FL has a great potential to solve the aforementioned issues in big data. It is proven in [13], [14] that FL can be used as an efficient solution to address the challenges of big data.

### A. Motivation and Our Contributions

Owing to the importance and practicality of FL and big data, there have been a number of surveys on these two topics. Several in-depth surveys on FL and its applications can be found in [15]–[22]. For example, the applications of FL for wireless and IoT networks are discussed in [15]–[17]. In particular, the work in [15] reviews FL-enabled solutions for resource allocations, reducing communications, mitigating privacy and security issues at mobile edge networks. A comprehensive survey of FL for IoT is presented in [16], where FL is used to enable IoT services such as data sharing, data offloading, attack detection, localization, and mobile crowdsensing, and IoT applications such as smart city, smart industry, smart healthcare, and smart transportation. Similar to [16], the work in [17] reviews the integration of FL with industrial IoT (IIoT) from three perspectives, including security, data management, and IIoT applications. There are several surveys that discuss the security, privacy, and threat issues of FL, for example, security and privacy issues in [18] and threats of FL in [19]. The use of FL is also reviewed for other applications such as healthcare [20], vehicular networks [21], and open banking [22]. Although these surveys on FL are extensive and can provide various lessons and future directives for the use of FL, they are very limited from the big data perspective, where big data services and applications are important.

In terms of big data, in-depth review articles can be found in many existing works. For example, the work in [23] provides a concise review on big data, including 1) relation of big data with enabling technologies such as IoT and cloud computing, 2) big data generation, acquisition, and storage, 2) big data applications such as social network, healthcare, and collective intelligence, and 4) a set of key issues and potential outlooks. There are also survey that focus on characteristics, roles, and applications of big data for specific field applications such as intelligent transportation systems, smart building, smart grids, IoT, and mobile networks [9], [24]–[27]. In particular, the work in [26] first presents a framework of big data in intelligent transportation systems, in which how big data is generated, collected, and processed is explained in detail, and important use cases such as accident analysis, traffic flow prediction, public transportation services, traffic route planning, asset management, and road control and management. The work in [27] surveys big data for IoT. In particular, this work focuses on reviewing big data technologies and approaches for different IoT domains, including healthcare, energy, transportation, building automation, smart cities, agriculture, industry, and military. This work also focuses on comparing big data

approaches for different IoT domains and summarizing key findings on the use of big data across IoT domains. A recent survey in [9] classifies big data deployment models and big data environments based on characteristics of the underlying network. First, four important features are presented from the deployment requirements, including resource management, security management, task management, and data management. Then, [9] presents big data environments such as cloud computing, decentralizing computing, and hybrid computing (i.e., a set of computing paradigms such as cloud computing, fog computing, and edge computing). Although the topic of big data has been well studied and reviewed in many surveys, the existing works only review artificial intelligence/machine learning for big data tasks, such as modeling, processing, and analytics, whereas the use of FL for big data has not been reviewed yet. The recent reviews on FL and big data are summarized in Table I.

Despite a number of important review articles on FL and big data, a comprehensive survey on the use of FL for big data has not been done yet. To fill this research gap, we aim to provide a concise survey on FL for big data. In particular, we focus on highlighting the opportunities that FL brings to big data services and applications. In summary, contributions and features offered by our work can be summarized as follows.

- **A preliminary to FL and big data:** Firstly, we present the fundamentals of FL and big data and discuss various motivations behind the use of FL for big data services and applications, such as mitigating security and privacy issues, reducing communication cost, handling the issue of data variety, easing the big data analysis, and improving the system scalability. We will detail this part in Section II.
- **FL for big data services:** FL plays an important role in big data services such as big data acquisition, big data storage, big data analytics, and big data privacy preservation. We review the use of FL for these big data services in detail in Section III.
- **FL for big data applications:** Next, we review FL-empowered big data in vertical domain applications, including smart city, smart healthcare, smart transportation, smart grid, online recommendation systems, social media, and several miscellaneous applications.
- **A review of practical projects on FL-big data:** Several practical projects on FL-big data are reviewed so as to drive real implementations of the FL-big data solutions.
- **Key challenges and promising directions on FL-big data:** Despite various advantages, the use of FL for big data also faces challenging problems such as bottleneck in communication efficiency, data heterogeneity, statistical heterogeneity, and privacy issues. These challenges will be explained in detail in Section VI along with potential directions of FL-big data.

### B. The Survey Organization

The organization of this paper is as follows. In Section II, we present the fundamental concepts of big data and FL, and the motivations behind their integration. Next, in Sections III

TABLE I  
SUMMARY OF RELATED REVIEW PAPERS ON FEDERATED LEARNING AND BIG DATA.

Ref	Contributions	Limitations
[15]	A preliminary to FL is presented, followed by a thorough review of FL-enabled solutions for privacy, resource allocation, and applications to mobile edge networks.	This survey only focuses on the use of FL at mobile edge networks.
[16]	A comprehensive survey of FL for IoT services and IoT applications is presented.	This survey only presents the use of FL for IoT services and applications.
[17]	A survey on fusion of FL and IIoT (e.g., background, data management, and IIoT applications) is presented.	This survey does not focus on the use of FL for big data.
[18], [19]	The sources of privacy/security/threat issues, types of attacks, unique features, and potential defensive solutions for FL.	These works are only limited to security/privacy/threat aspects of FL.
[20]–[22]	The use of FL for particular applications, e.g., healthcare, vehicular IoT, and open bank are presented.	These studies do not focus on the role of FL for big data services and applications.
[23]	A concise review on background, key techniques, and important applications of big data	This survey was done before the invention of FL, and thus lacks the role of FL for big data services and applications.
[26]	A comprehensive survey of big data for intelligent transportation systems, including big data characteristics, applications, and platforms.	This survey only focuses on machine learning for big data modeling and analytics, while the use of FL is ignored.
[27]	This survey reviews the state-of-the-art big data technologies for different IoT domains and presents key findings when big data is used across IoT domains.	The use of FL for big data services and applications is not presented.
[9]	This work presents two important aspects of big data: deployment requirements and network environments	The use of AI and FL for big data services and applications is not presented.

and IV, we present the use of FL for various big data services and applications, respectively. After that, we review a number of projects in Section V to show the practicality of the FL for big data. In Section VI, we then discuss some research challenges and highlight promising future directions to further drive the use of FL for big data. Finally, we conclude the paper in Section VII.

## II. FEDERATED LEARNING AND BIG DATA: AN OVERVIEW

In this section, we discover the fundamental concepts of FL and big data. The motivations of their integration is given correspondingly.

### A. Federated Learning

Recognized as a new learning paradigm where statistical models are trained on multiple edge devices in a distributed network without sharing their training data, FL provides AI tools to train a model collaboratively by using a federated dataset of secure sources. With the increasing computational power and storage capacity of smart devices (e.g., mobile phones, wearable devices, and autonomous vehicles), storing data and processing data are executed at the edge locally in distributed networks, especially in this era when the information privacy is priority. Transmitting raw data of mobile devices over wireless communications is so sensitive to cyberattacks, which in turn forces the rapid growth of FL. Here we discover the operation of FL, give some common FL categories, and discuss its core challenges. A general FL architecture with processing flow is presented in Fig. 1.

**Taxonomy of FL** Based on the examination of existing FL frameworks, we categorize FL regarding training data property and networking topology.

**Data property:** This category can be divided into three subclasses, including horizontal FL (HFL), vertical FL (VFL), and federated transfer learning (FTL), based on the properties of training data [28].

- **Horizontal FL:** In HFL frameworks, the edge devices local train statistical models using their own datasets having the same feature space, but different sample space. Due to this property of input training data, all devices should utilize a same AI model (e.g., support vector machine and deep neural networks). For an example of HFL applications, the hospitals collect the health data of different patients with the electronic medical records having the same attributes.
- **Vertical FL:** Different from HFL, VFL is introduced for the scenarios, in which the local datasets share the same samples but differ in features. To collect overlapping data samples in different edge devices, an entity alignment mechanism is utilized. For an example of VFL scenarios, different hospitals collect the health data of the same patient but differ in attributes, such as diabetic retinopathy in hospital A, cardiovascular disease in hospital B, and diabetic nephropathy in hospital C.
- **Federated transfer learning:** FTL is proposed for the unrelated learning scenarios, in which the local datasets totally differ in sample space and feature space. In this framework, the knowledge of local models is transferred across domains by the FTL server. For example, in the healthcare domains, disease diagnosis can be collaborated by multiple hospitals in different countries, where the electronic medical records of various tests are collected from different patients.

**Network topology:** Based on the network topology of model communications, this category can be divided into two

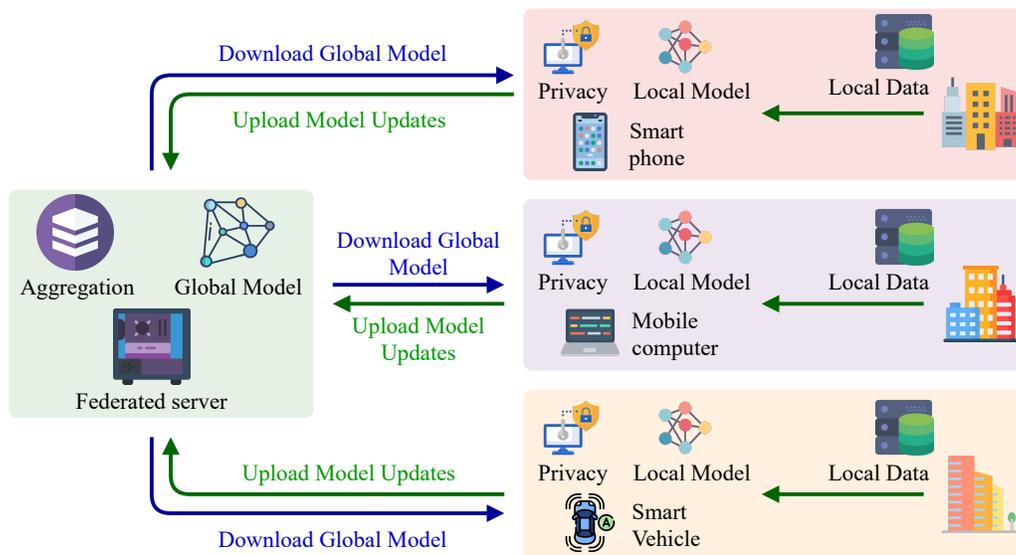


Fig. 1. General architecture of FL with processing flow.

sub-classes: centralized FL and decentralized FL [29].

- In a centralized FL system, all devices transmit the trained parameters of local models to the FL server for updating the global model with a parameter aggregation algorithm. Afterwards the sever delivers the computed global model to all clients in a network for the next training iteration.
- In a decentralized FL system, all devices are connected over peer-to-peer communications. A device can transmit the model parameters, locally trained on its own dataset, to neighbors and receive their model updates to aggregate parameters.

### B. Big Data

Nowadays, a massive data volume are generated daily with unprecedentedly increasing rate from heterogeneous sources. This is due to the emergence of the internet of things, the proliferation of the cloud computing, and the spread of smart devices. In fundamental, big data refers to massive datasets with heterogeneous formats: structured, unstructured, and semi-structured data, which has three key characteristics: large volume with the incessant generation from millions of devices and applications, high velocity with the real-time data acquisition and processing, and high variety with various source and multiple formats. **Taxonomy of big data** Here we present a taxonomy of big data in various perspectives, including data domain, storage infrastructure, compute infrastructure, and AI-based data processing [30]. This taxonomy is illustrated in Fig. 2.

**Data domain:** Various domains, in which big data has being arisen in the last decade, can be determined based on the time span of processing and the degree of data structure. For the time span of data processing analysis, data domain can be classified into three groups: batch (including bioinformatics, geosciences, forensics, and other large-scale sciences), near real-time (including retail and sensory data), and real-time (including finance, social networking, and network security) [31].

We can also classify various domains into three degrees of structure: structured data - containing relational databases and spreadsheets (including retail, finance, bioinformatics, and geosciences), unstructured data - containing raw signals from nature or data without pre-defined structural model (including voice, image, video, and sensory data), and semi-structured data - as the hybrid form of structured and unstructured data (including web logs, email, and documents). Notably, relying on specific applications, the visual media domain with high-dimensional unstructured data can spread from batch to real-time of time span [32].

**Storage infrastructure:** Numerous big data storage solutions have been introduced to engineers and developer when they build services and applications running with large datasets. In Fig. 2, we can find the taxonomy of different types of databases being used for big data storage. The storage platforms in the relational structured query language (SQL) group are with some features: speed over scale, large capacity for vertical scaling, good consistency over availability, and easy deployment. Inspired by relational SQL, newSQL offers higher performance and scalability, thus enables high throughput online transaction processing requests while maintaining high-level language query capability. In the nonSQL group, we have several database techniques for specific purposes, such as document oriented, graph oriented, key-value storage, and big table inspired. Although numerous database platforms have been introduced for diversified services and applications with different data types, they should comply with some standards, for example ACID (atomicity-consistency-isolation-durability) and BASE (basically available-soft state-eventually consistent).

**Compute infrastructure:** Based on the time span metric, computing paradigms for big data can be classified into two primary categories: batch mode and real-time/near real-time mode. For the batch mode with high processing latency, MapReduce and Bulk synchronous parallel are recommended for large parallel and distributed systems. Whereas low-latency

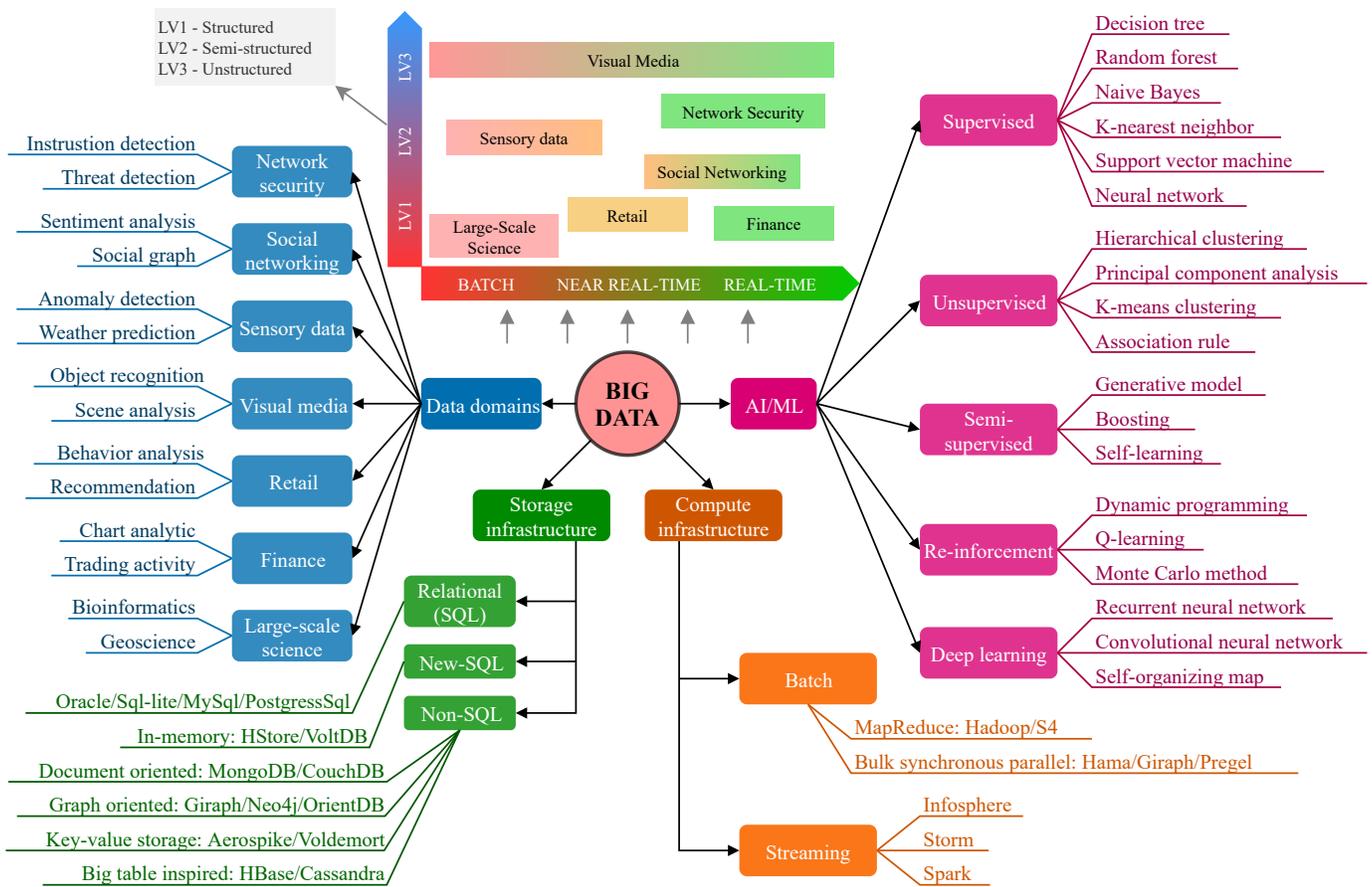


Fig. 2. Taxonomy of big data technology.

computing frameworks for streaming demand immediate response to every incoming event. Some popular frameworks in this group are Infosphere, Apache Storm, and Apache Spark.

**AI-based data processing:** Artificial intelligence (AI) techniques and machine learning (ML) algorithms allows analyzing and processing data in automatic and scalable ways, in which the high-level information and meaningful context can be derived from raw data. ML allows computers to learn complicated patterns automatically and make decisions of new incoming data events [33]. Depending on data utilization of model learning, these algorithms can be categorized into four classes: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. Recently, deep learning has emerged with recurrent neural network, convolutional neural network, and self-organizing map, which has achieved remarkable success in various domains due to the great capability of dealing with large noisy-messy-confusing datasets.

### C. Motivations of the Integration

With the unprecedented growth of IoT services and applications, traditional AI/ML techniques based on centralized computing platforms have exposed several issues: security, privacy, variety, communication, analysis, and scalability. With the ability to train statistical models on own datasets of

edge devices locally and then transfer model parameters to the centralized server for global model aggregation, FL is recognized as a promising solution to solve these above-mentioned problems of big data. Along with the integration of FL and big data, the main motivations behind such kind of this integration are summarized as follows.

- **Security and data privacy:** To extract meaningful information and capture patterns of the data generated by edge devices, traditional AI and ML algorithms are usually adopted in the centralized server to train statistical learning models. To this end, the data should be transferred from devices to the server for training a model, which in turn exposes some security and privacy risks over cyberattacks. Instead of uploading data to the server, FL allows devices to train models with their datasets locally and transfer model parameters to the server for aggregation.
- **Communication cost:** A massive volume of raw sensory data and low-level informative data of clients should be transferred to the centralized server for processing, which consequently suffers a considerable communication cost, especially with high-dimensional unstructured data like images and videos. With FL, the communication cost reduces significantly owing to transferring trained statistical models to the FL server instead of the entire data.

- **Data variety:** In several integrated services and applications, the multimodal sensory data acquired from various sources has different types and structures (structured data vs unstructured data and time series data vs high-dimensional data), which in turn makes difficult to manage in the big data server deploying a specific storage platform. Besides, each data type characterized by various attributes will demand a specific computing infrastructure to achieve best performance of target services and applications. With FL, each devices and clients may take into consideration a specific data type locally, thus reducing the complexity of local data management at the FL server.
- **Analysis:** In many centralized computing systems, learning classification and regression models with the multimodality data can be accomplished using advanced methods (e.g., data-based, feature-based, and decision-based fusing mechanism), but its processing framework will be much more complicated. Whereas FL can offer a universal DL model to different data modalities of a same learning task.
- **Scalability:** In conventional centralized systems, the scalability of storage and computing infrastructures at the server is problematic because of the complex nature of big data with 5Vs. In the FL platforms, the server only performs an aggregation function with multiple locally trained models to modernize the global model, hence system scalability is capable of performing at edge devices. Intuitively, the scalability of FL is more flexible and convenient than that of centralized learning in the server.

### III. FEDERATED LEARNING FOR FOR BIG DATA SERVICES

FL can play a very crucial role in big data services such as acquisition of big data, storage of big data, big data analytics, and privacy preservation of big data. In this section, the role of FL for big data services is discussed along with recent state-of-the-art. Fig. 3 illustrates FL for big data services.

#### A. Federated Learning for Big Data Acquisition

The exponential growth of AI, ML, and Big data analytics has a major impact on a wide range of applications, including health care system, smart grid, smart cities, intelligent transport systems, and so on. Due to the rapid growth of the applications, large volumes of data are being generated at an exponential rate, encouraging the applications to generate the data at a rapid pace. Moreover, some companies and organisations employ advanced artificial intelligence (AI) approaches and big data analytics to quickly identify consumer behaviour in order to increase productivity, sales, and profits. Data sharing is a key problem in AL and ML to fulfil these requirements. Acquiring public data and performing data analytics, is not a perfect solution since public data has constraints such as poor data quality, inconsistency, and unstructured. Most of the companies prefer not to share their data for a variety of reasons and concerns, such data privacy, data sharing rules, and a lack of information. Some times the companies may occasionally share information with third parties in order to

facilitate growth and mutual data cooperation agreements. The primary goal of FL is to train ML algorithms across other distributed devices using its own private and local data. The locally trained models are first transmitted to the server, and then all of the shared models are combined with other updates to build a better global model, which is then shared with all devices. The main advantage of FL is that the training data is held on the device, with no individual updates sent to the cloud.

1) *Federated Learning for Secure Big Data Collection:* The most important aspect of FL is that it helps in training shared statistical on the basis of decentralized devices or servers in a local data set. Although the data scientists and researchers use similar model to train the data, the use of FL eliminates the need to upload private data to the cloud or the exchanging of data with other members in the scientific community. In case of traditional machine learning approaches wherein the data resides on a single server, FL reduces the challenges associated with achievement of data security and privacy by maintaining the storage of local data. FL has been extremely successful in handling the various challenges relevant to user privacy protection thereby achieving data security. In FL, the data provisioned at the end-user systems is decoupled and the machine learning models are aggregated ensuring involvement of network parameters in the implementation of deep learning at the centralized server. The primary objective of FL is to understand and learn the global model without compromising with the data privacy. IT has unique abilities to achieve data privacy while training the data in a dataset. In case of "anonymized" dataset stored in a server, there are possibilities of client privacy getting compromised if the dataset is linked to other datasets in the framework. On the contrary, in case of FL the information transmitted using FL has minimum updates which contributes in improvement of accuracy in a machine learning model. One of the studies conducted by [14] used a Verifiable FL (VFL) to enable preservation of privacy for Big Data in an Industrial IoT setting. Although the traditional deep learning techniques have been quite successful, but in case of privacy preservation, the traditional data gathering based centralized learning fails to achieve privacy and security in industrial scenarios for the training of data. In a VFL setup, instead of gradient aggregation, Lagrange interpolation is used to set the interpolation points which helps to validate the accuracy of the aggregated gradients. The verification overhead of VFL thus remains constant inspite the number of participants which help in getting optimum level of accuracy. Another example of FL implementation in big data security is found in [34]. It is a known fact that the credit card transaction datasets are generally skewed and the number of data relevant to fraudulent activities are quite less. Hence it becomes extremely difficult to understand and analyse patterns of frauds and later detect actual frauds. The study proposed a FL for Fraud Detection (FFD) framework which enabled the banks to learn the fraud detection model by training the data distributed in their localised database. The FDS is constructed by the aggregation of local computational updates of the fraud detection model. The banks thus can achieve data privacy without having to share the data or sensitive information of

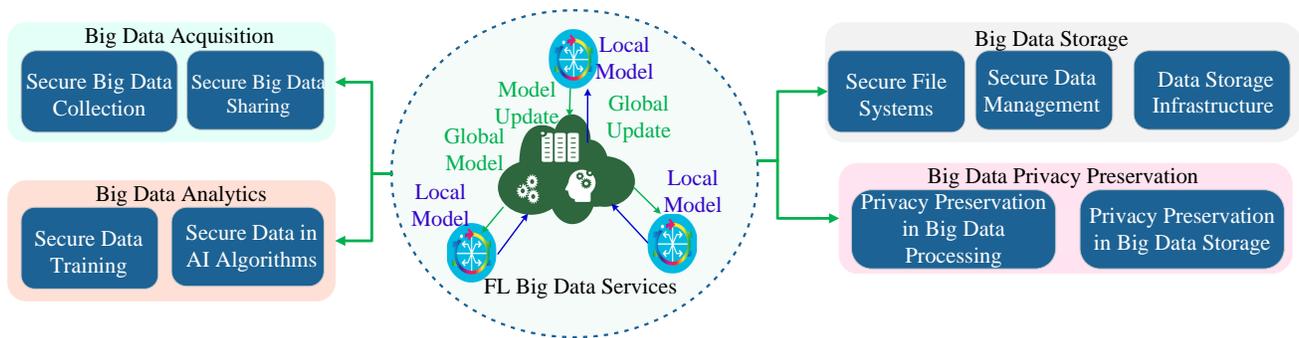


Fig. 3. FL for big data services.

the card holders.

#### 2) *Federated Learning for Secure Big Data Sharing:*

Data sharing in big data applications helps in transferring data in a shared network in order to help end users access and use various applications. FL eliminates the process of sharing raw data and takes an alternative approach of sharing the learned results for preserving privacy and ensuring low latency of applications. The study in [35] discusses a integrated data sharing model implemented for industrial IoT systems. The data owners and requestors in the study are able to achieve fast and secure exchange of data within the involved decentralized parties. The FL scheme also addresses considers resource constraint issues of IoT systems. The study in [36] implemented a framework for a scenario based on industrial IoT, in which FL is used to achieve optimal level of security. In this work, various factories collaborate with the tensor mining to share their data which are encrypted using homomorphic encryption technique in the centralized server. FL can also be implemented to achieve distributed data sharing in vehicular networks. Towards this direction, the authors in [37] presented an asynchronous federated data sharing architecture for internet of vehicles. In this case, the vehicle is an FL client used to share data with the aggregation server located at macro base station (MBS). The vehicles with various service requests pertinent to prediction of traffic or path detection post their requests to the MBS. The MBS implements computing tasks to resolve these requests using an actor-critic reinforcement learning framework. Thus, the system segregates good and bad nodes, enables secured and intelligent data sharing, decision making, thereby ensuring cost optimization.

#### B. *Federated Learning Big Data Storage*

In this subsection, we discuss how FL can be used for storage of big data.

1) *Federated Learning for Secure File Systems:* In traditional FL, due to some reasons, if the central server becomes unavailable, the entire training process will be disrupted. The server should also have high-bandwidth and reliable communication links with the agents to transfer big data. Also, all the agents should trust the server. To overcome the aforementioned issues, a decentralized approach, where the model relies on the own resources of the node, can be used as an alternative to perform the training. Pappas et al. [38]

proposed a decentralized FL framework, called, interplanetary learning system, that is inspired by interplanetary file system. The proposed framework allows mobile agents to collaborate in the model's training that doesn't rely on any central server.

2) *Federated Learning for Secure Data Management:* Due to the latest industrial revolution (Industry 4.0), the user equipments are distributed and widespread. Conventional network infrastructures are unsuitable for Industry 4.0 applications due to communication delays and transmission media costs. One of the major enablers of Industry 4.0 is Industrial IoT [39]. Several Industrial IoT applications like driverless cars, intelligent robots, smart medical systems, and smart grid are based on wireless networks that generate big data [40]. Many researchers are working on efficient storage, usage, and management of big data generated from Industrial IoT [41]. The big data generated from Industrial IoT often contains private and sensitive information that should not be exposed as it is vulnerable to attacks from malware, heterogeneous equipments or heterogeneous networks [42]. FL can address the aforementioned challenges in the management of big data generated from Industrial IoT devices as it preserves privacy of the users' data by not allowing the global ML model to have access to the sensitive information [43]. FL can also maintain the heterogeneity of the data, hence reducing the deviation of training in ML model. To reduce the training rate of ML model on big data from Industrial IoT and to reduce the model aggregation's communication cost, Zhang et al. [44] proposed a framework based on FL assisted by deep reinforcement learning. In a similar work, Kim Sungwook [45] proposed an incentive mechanism for attracting several owners of data for joining in the process of FL to preserve privacy and efficient management of data. The proposed approach adopts two concepts: mechanism design for designing incentives to achieve the objectives, and differential privacy for privacy preservation. Data management in real time is essential in remote monitoring of patients. To take the decisions in a timely manner, the processed data from the patients big data has to be provided to the care providers. Several issues, such as patient mobility and computational complexity may arise in processing the data in real time. FL can be used effectively in collaboration with technologies such as cloud computing and fog computing, ML based approaches to solve these issues [46].

For handling the challenges and open issues in big data

streams generated through IoT devices in smart cities in terms of processing of big data and networking, Imteaj et al. [47] proposed a distributed sensing scheme that paves the way for FL for distributed decision making. In this scheme, a device can be identified by a token that can activate the user's devices that are distributed for sending the data to the cloud dynamically and the data can be stored in the cloud in a proper format. The proposed approach ensures data collection remotely using the end-user devices that are available and hence the cost incurred in installing new sensors for IoT applications can be reduced.

### 3) *Federated Learning for Data Storage Infrastructure:*

The standard machine learning techniques require training of the data in one machine or a data center. FL enables devices to learn from a shared prediction model ensuring that all the training data is stored in one device. It decouples the ability to implement machine learning based on the need to store data in the cloud. As an example, when a device downloads the current model, it improvises the same by learning from the data in the device and then summarizes the changes in the data as specific update to the model, which alone is sent to the cloud. In this process, encrypted communication technique is used wherein it gets immediately averaged with other updates that initiates improvement of the shared model. Thus all the training data remains on the local device and no updates get stored in the cloud. Thus it eliminates the need to store data in the cloud and also incorporates the use of "Secure Aggregation Protocol" using cryptographic techniques such that a coordinating sever is able to decrypt only the average update when hundreds and thousands of users have participated. It does not allow any individual update to be inspected unless averaging is done [48], [49].

### C. *Federated Learning for Big Data Analytics*

The growth in IoT based applications is resulting in generation of massive sensing data at a rapid pace. The generated data may contain important information that can be analyzed by understanding the patterns existing in the raw data. ML and allied technologies can be used effectively for big data analytics. Big data analytics require the participant sources to send their personal and sensitive data to a central sever to perform ML services, that may result in privacy concerns of the data sources such as mobile users. FL can be used efficiently to provide privacy reservation while providing big data analysis [13]. In this subsection, the application of FL for secure data training and securing the data in AI algorithms is discussed.

1) *Federated Learning for Secure Data Training:* The current approach of performing big data analytics in the cloud has concerns related to network cost and data privacy. FL can address these challenges by training the data in local devices and updating the global model with the local parameters without sharing the raw data. The recent state of the art FL for secured training of big data is presented below.

Xu et al. [50] have proposed a novel framework based on FL, named FL-Pruning, Quantization and Selective Updating (PQSU) to address the network connectivity issues and limited

computational resources for performing big data analytics in IoT devices that slows down the training process. The proposed method consists of 3 phases, namely, structured pruning, quantization of weights and updating selective, that collaborate to reduce the storage, communication and computation costs, thereby fastening up of the training process of big data from IoT devices. In a similar work, to address the problem of leaking of sensitive information of the participants to the untrusted servers in big data analytics, Zhang et al. [13] proposed PEFL, a privacy-enhanced FL approach for protecting gradients in a server which is untrusted. This is achieved by encrypting the local gradients of the participants by using paillier homomorphic cryptosystem. For reducing the cryptosystem's costs of communication of the the authors have used a distributed selective stochastic gradient descent approach for achieving distributed encryption during the training phase. The encrypted gradients enable secured aggregation in server. Hence, the untrusted server can learn only the statistics of the participants that are aggregated, while preserving the privacy of the individuals. In another work, Doku et al. [51] proposed a FL and blockchain based model to retrieve relevant data from big data by using proof of common interest mechanism to segregate relevant data from big data. This data is then trained on a federated model on the edge devices. A shared model is then generated by aggregating the model along with other models in the network, which is then stored in the blockchain. The network members can download the aggregated model that can be used to provide edge intelligence to the users.

In a smart grid environment, massive data is generated from the households and industries. Smart metering systems in smart grid have to perform data analytics quite frequently (may be once in every 15 minutes) to understand the patterns in consumption of electricity by the clients, that can enable coping up with the usage of electricity during peak times, seamless integrated of renewable sources of energy through coordinate consumption of electricity. Even the retailers can be benefited by big data analytics performed on the data generated by smart meters to have more transparency of consumer's behavior in consuming electricity, through which personalized services can be provided to them. In the retail market, several retailers own the data from smart meters. They may not share the data to the central cloud for big data analytics due to privacy concerns. Hence, the identification of the characteristics of the consumers is difficult as full dataset can be accessed by the global ML models. To address this issue, Wang et al. [52] proposed a FL based approach to train the learning model on the local data residing at diversified clients without sharing the raw data to the global ML model, thereby preserving the privacy of the data. In this work, principal component analysis is executed to extract the significant features from the consumer data. Artificial neural networks based ML algorithms are trained on the consumer data using FL approach for big data analytics. In a similar work, Zhai et al. [53] proposed an FL framework, constrained by delay deadline that overcomes long delays due to training of big data generated in smart grids.

## 2) *Federated Learning for Secure Data in AI Algorithms:*

Due to strict regulations and protections of data security and privacy, traditional ML algorithms that get trained on the datasets that are located on central servers are facing significant challenges. Due to these reasons, conventional AI and ML based approaches are becoming impractical in several data-sensitive and mission critical scenarios, like health, government, finance sectors, etc. FL has received wide attention recently because of its ability to preserve privacy of users as it doesn't send the raw data of the users to update the parameters of the global model. But, FL can still be vulnerable to several privacy and security threats as the attackers can use the shared gradients to derive the privacy of the participants. To solve this problem, several researchers have proposed various privacy preserving and secure FL approaches. Some of the recent state of the art in this direction are discussed below.

Xu et al. [54] have proposed VerifyNet model that ensures security and verifiability of FL algorithms that train on big data. In this work, firstly, a double-masking protocol is used to ensure that the local gradients of the users are confidential during the FL. Later, to prove the correctness of the aggregated results to the users, the cloud server has to provide the proof. In a similar work, So et al. [55] proposed a single-server Byzantine-resilient secure aggregation framework for a secured FL. The proposed approach is based on secure model aggregation, verifiable outlier detection, and integrated stochastic quantization to ensure privacy, convergence, and Byzantine resilience at the same time. In another interesting work, Li et al. [56] proposed a novel Knowledge Federation framework, to preserve to preserve privacy and securing the training data, utilize all the data resources that are scattered across several units of organizations.

## D. *Federated Learning for Big Data Privacy Preservation*

1) *Federated Learning for Privacy Preservation in Big Data Processing:* FL acts as a promising solution catering to solve problems associated with data islands, breaking of data barriers. It ensures protection of data privacy and security of big data. In a Distributed IoT framework, it is extremely essential for the users to collaboratively train the classification or regression model to achieve data prediction ensuring optimum level of privacy. Instead of preserving privacy in outsourced training, the users train the data locally using FL rather than submitting the same to the central server. The federated center remains responsible for only aggregation of gradient information which are uploaded by users during the distribution of global training model [12], [57]–[59].

2) *Federated Learning for Privacy Preservation in Big Data Storage:* In the present day and age, there exists multiple devices that help to connect with one another generating massive amount of data commonly termed as "Big Data". These devices include wearables, autonomous vehicles, mobile devices which require ever growing computational power and also has privacy concerns while storing or processing of the same. This pushes the storage mechanisms towards edge computing leaving cloud. Edge computing thus acts as a new paradigm used for the deployment of computationally intense

applications and operation of the same. FL in such scenario is an emerging technique that enables privacy preservation while training the Deep Neural Network Model when data gets generated from multiple clients. It is a combination of distributed machine learning, cryptography, security and other incentive based techniques following the basic principles of economy and game theory. In FL, the machine learning based framework consists of many clients who collaboratively train a model with the support of a centralized server ensuring that the training data is decentralized. The machine learning algorithms in such case gets trained at various local datasets located in the local edge nodes. The FL based system keeps the raw data distributed in the client devices instead of aggregating the same to the centralized cloud data center to perform training. The shared model is trained on the server by the aggregating the computed updates present locally [12], [57]–[59].

## E. *Summary*

Due to the rapid growth of the applications, large volumes of data are being generated at an exponential rate, encouraging the applications to generate the data at a rapid pace. From the discussion above, it can be summarized that FL has a tremendous potential in addressing several challenges faced by big data services such as Big data acquisition, Big data storage, Big data analytics, Big data privacy preservation. However, several challenges such as many companies prefer not to share their data for a variety of reasons and concerns, such data privacy, data sharing rules, and a lack of information that can be solved using FL. Performing big data analytics in the cloud has concerns related to network cost and data privacy. Poor latency for data exchange among decentralized multiple parties. Leaking of sensitive information of the participants to the untrusted servers are addressed to realize the full potential of FL for big services. Table II summarizes the benefits and challenges of FL for big data services.

## IV. FEDERATED LEARNING ENABLED BIG DATA APPLICATIONS

In this section, the advantages of FL for several big data applications are discussed along with recent state of the art literature.

### A. *Federated learning enabled big data in smart city*

A smart city improves the quality of life of the urban citizens by optimizing the functions of the cities and promotion of economic growth by making use of data analysis and smart technologies like IoT. IoT sensors are placed throughout the smart cities to collect information regarding traffic, air quality, condition of the roads, garbage, etc. These sensors generate large volumes of data continuously and at a fast pace. Big data analytics along with ML based algorithms can effectively use these data to gain the insights. The insights gained can be used to effectively manage the services, resources, and assets in cities. Big data analytics can provide solutions to some of the important issues like healthcare, IoT, communication, transportation, garbage collection, disaster management, etc

TABLE II  
BENEFITS AND CHALLENGES OF FL FOR BIG DATA SERVICES.

Services	Existing Challenges Faced	Benefits of using FL
Big Data Acquisition	<ol style="list-style-type: none"> <li>1) Most of the companies prefer not to share their data for a variety of reasons and concerns, such data privacy, data sharing rules, and a lack of information.</li> <li>2) Data security and privacy for maintaining the storage of local data.</li> <li>3) Difficult to understand and analyse data patterns of fraud transactions and later detect actual frauds.</li> <li>4) Poor latency time for data exchange among decentralized multiple parties.</li> </ol>	<ol style="list-style-type: none"> <li>1) Training data is held on the device, with no individual updates sent to the cloud.</li> <li>2) To understand and learn the global model without compromising with the data privacy.</li> <li>3) Achieve data privacy without sharing the data or sensitive information of the users.</li> <li>4) Achieve secure and fast data exchange among decentralized multiple parties.</li> <li>5) FL with Blockchain build immutable data blocks which provides transparency and improves data ownership without the need of any central authority.</li> </ol>
Big Data Storage	<ol style="list-style-type: none"> <li>1) The central server becomes unavailable, the entire training process will be disrupted.</li> <li>2) The server have poor bandwidth and, unreliable communication links with the agents to transfer big data.</li> <li>3) Poor data storage infrastructure.</li> <li>4) Poor latency time for data exchange among decentralized multiple parties.</li> </ol>	<ol style="list-style-type: none"> <li>1) FL provides a decentralized approach, where the model relies on the own resources of the node, can be used as an alternative to perform the training.</li> <li>2) FL allows mobile agents to collaborate in the models training that does not rely on any central server.</li> <li>3) Maintains the heterogeneity of the data, hence reducing the deviation of training in ML model.</li> <li>4) Decouples the ability to implement machine learning based on the need to store data in the cloud.</li> </ol>
Big Data Analytics	<ol style="list-style-type: none"> <li>1) Big data analytics require the participant sources to send their personal and sensitive data to a central sever to perform ML services,that may result in privacy concerns of the data sources such as mobile users.</li> <li>2) Secure Data Training.</li> <li>3) Performing big data analytics in the cloud has concerns related to network cost and data privacy.</li> <li>4) Leaking of sensitive information of the participants to the untrusted servers.</li> <li>5) Conventional AI and ML based approaches are becoming impractical in several data-sensitive and mission critical scenarios, like health, government, finance sectors,etc.</li> </ol>	<ol style="list-style-type: none"> <li>1) FL provides privacy reservation while providing big data analysis.</li> <li>2) Addresses network connectivity issues and limited computational resources for performing big data analytics in IoT devices that slows down the training process.</li> <li>3) Fastening up of the training process of big data from IoT devices.</li> <li>4) The network members can download the aggregated model that can be used to provide edge intelligence to the users.</li> <li>5) Ensures security and verifiability of FL algorithms that train on big data.</li> </ol>
Big Data Privacy Preservation	<ol style="list-style-type: none"> <li>1) Privacy Preservation in Big Data Processing.</li> <li>2) Privacy Preservation in Big Data Storage.</li> <li>3) Preserving privacy of sensitive information of users.</li> <li>4) Poor security and slow data exchange among decentralized multiple parties.</li> </ol>	<ol style="list-style-type: none"> <li>1) Solve problems associated with data islands, breaking of data barriers.</li> <li>2) Ensures protection of data privacy and security of big data.</li> <li>3) Keeps the raw data distributed in the client devices instead of aggregating the same to the centralized cloud data center to perform training.</li> <li>4) The shared model is trained on the server by the aggregating the computed updates present locally.</li> </ol>

in smart cities [60]–[62]. However, most of the data generated from the sensors in smart cities is sensitive in nature which involves private data of its citizens [63]. So, the need of the hour is to preserve the private data generated in smart cities and efficient handling of network resources to handle the big data. Traditional ML based methods cannot handle these issues that can arise in the smart cities. FL, through its inherent characteristics, is a potential solution that can address the aforementioned issues in smart cities as depicted in Fig. 4 [64], [65]. Recent studies on FL used for addressing the challenges of big data from smart cities are discussed below.

Prediction of traffic flow plays a significant role in management of intelligent transportation in smart cities. By analyzing the patterns from the big data generated related to the traffic flow and making predictions can help in realizing several benefits like reducing traffic congestion, accidents, air pollution. However, as the driver's information is tightly coupled with the location of the vehicle, the data collected from the vehicles may threaten the privacy of the driver. The attacker/malicious user can get hold of personal details of the driver like health condition, habits, religion, income, etc. based on the places visited frequently by the driver. Motivated by these issues, Qolomany et al. [66] have proposed to use a FL model to preserve the privacy of the user. The authors have proposed to use a deep LSTM model to train the data locally. The parameters for the LSTM model are chosen by particle swarm

optimization algorithm. Each client will only compute and share the update with the global model residing at the server. In this way the private and sensitive data will not be transferred to the central storage, preserving privacy of the drivers.

Safety monitoring system is a very important application (like monitoring of fire hazards) in a smart city. Building models for object detection/recognition on large datasets that are stored centrally is a very challenging task due to high cost involved in transmitting video/image data and privacy issues. To overcome these challenges, Liu et al. [67] proposed a machine learning platform, namely, Fedvision, that leverages FL in computer vision related applications. Fedvision is deployed with the collaboration between Extreme Vision and WeBank for smart city applications that can be used by the customers in developing safety monitoring systems. One of the challenges in smart city applications is to process and get insights from large volumes of unlabelled data being generated at a very rapid rate. Also, the privacy of the data has to be preserved. To overcome these challenges, Albaseer et al. [68] proposed a FL based semi-supervised method for real-time analytics at the edge networks, FedSem, for smart city applications.

One of the primary goals of a smart city is to build a smarter and better healthcare infrastructure. Training large volumes of medical data in a central server involves several challenges such as preservation of privacy of patients, heterogeneous nature of data from different sources, etc. To address these

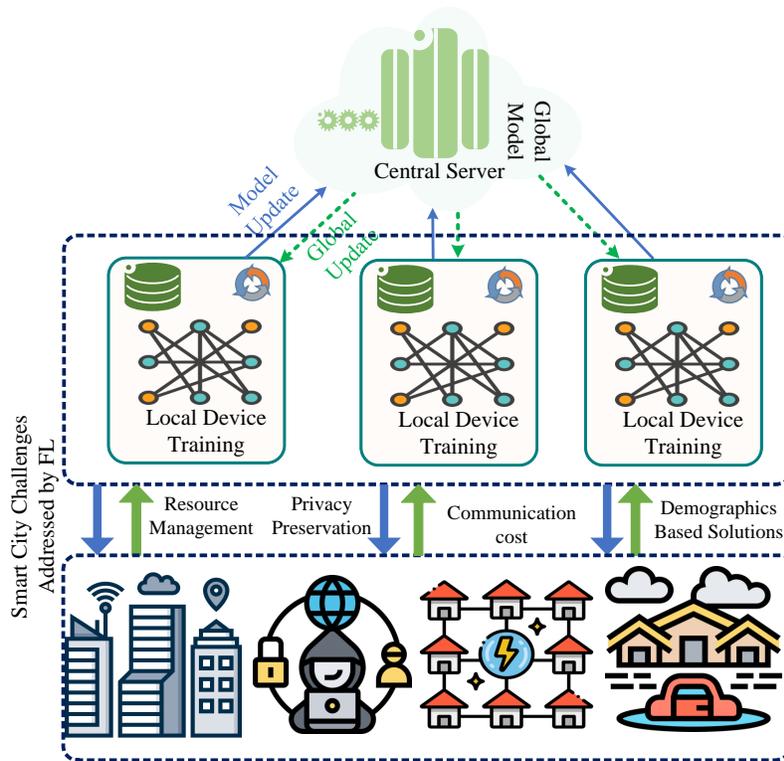


Fig. 4. Federated Learning Enabled Big Data in Smart City.

challenges in large volumes of medical data, Thwal et al. [69] proposed a FL paradigm based on deep learning.

When facing with crisis like COVID-19, city digital twin can face several challenges due to its reliance on high quality and long term data. To address these challenges, Pang et al. [70] proposed an FL based city digital twin to efficiently accumulate insights for heterogeneous sources of data. The global model is trained iteratively at different city digital twins till the global model uncovers the patterns between the response plans and trends in infection. In this way, the proposed model can be used to manage the crisis in a city.

Unmanned Aerial Vehicles (UAVs) play an important role in smart cities to monitor traffic, air pollution, movement of citizens during pandemics, and disaster management [71]. However, the data acquired by the UAVs may be very sensitive and the privacy preservation of the data acquired is of paramount importance. Hence, UAV integrated with FL can preserve the privacy of the sensitive data, address several concerns like resource management, latency issues faced by UAVs. For forecasting and monitoring of air quality index in a smart city from spatial-temporal perspective, Liu et al. [72] proposed an aerial-ground air quality sensing framework with UAVs swarms based on FL that addresses the privacy concerns of sharing the air pollution related data. In a similar work, the authors in [73] leveraged FL for UAVs in smart city monitoring for truthful reporting from UAVs and also to address heterogeneity of data generated from multiple UAVs.

#### B. Federated Learning enabled Big Data in Smart Healthcare

AI/ML techniques integrated with IoT have been extensively used over the past two decades in healthcare applications to assist the medical practitioners in diagnosing the diseases, drug discovery, remote patient monitoring, etc. Traditional ML approaches face several issues like exposing the private and sensitive information of the hospital/patients, issues in sharing huge volumes of data, etc. With FL, the computation model itself will be executed at the data source; hence, FL has a huge potential in addressing the aforementioned issues in smart healthcare applications as depicted in Fig. 5. FL can enable large-scale precision medicine respecting individual privacy concerns [74]. The recent works on applications of FL for big data related to smart healthcare is presented in the rest of the subsection.

Hakak et al. [75] have proposed a edge assisted framework with FL for privacy preservation of sensitive healthcare data acquired through wearable devices. Apart from privacy preservation, the proposed framework helps in optimizing cloud resources when dealing with huge volumes of healthcare data generated by millions of wearable devices. In a similar work, Qayyam et al. [76] proposed a edge assisted framework for automatic diagnosis of COVID-19 using clustered FL. The proposed framework can help in performing analytics for the large volumes of ultrasound and x-ray data related to COVID-19 patients at the FL enabled edge. Silva et al. [77] have proposed a framework based on FL to train the ML models on databanks that contain brain images. The proposed approach is mainly aimed at privacy preservation of the sensitive information of the patients. Wu et al. [78]

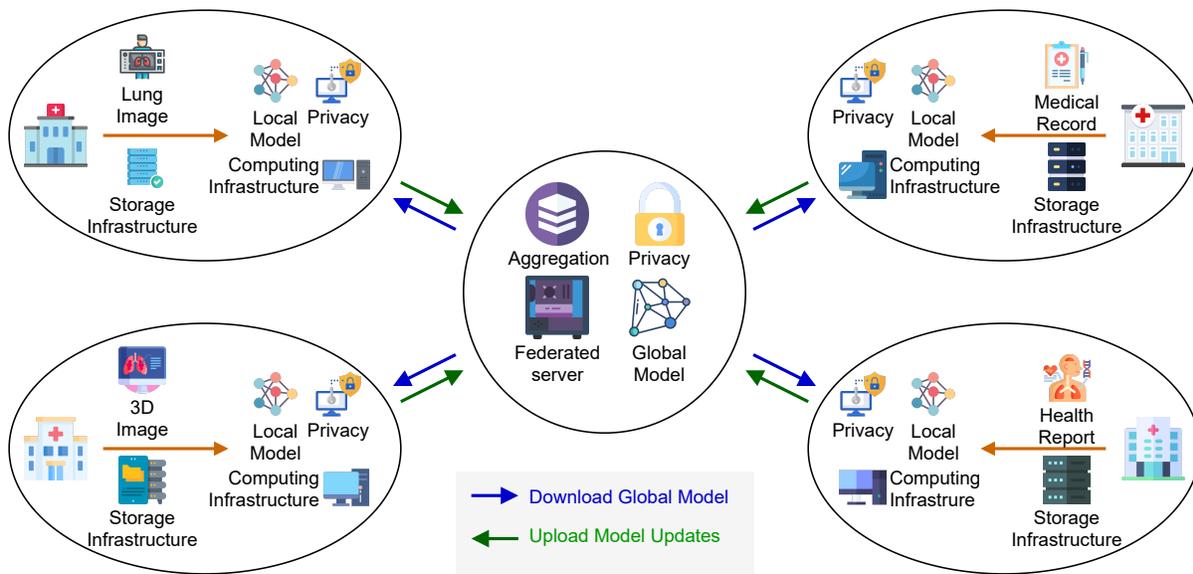


Fig. 5. Integration of FL and big data for medical and healthcare services.

have proposed a novel framework, FedHome, that is based on FL for in-home monitoring of health. The authors aim to address the existing challenges of privacy preservation and transferring of large volumes of data that exist using the ML algorithms are used on the in-home patients' data in a central server. To get diverse and large datasets to train a deep learning model, multi institutional collaboration is need of the hour to generate diverse and large quantity of medical data. How to enable collaboration among multiple institutions without exposing the sensitive information of the patients is a challenge. To address this issue, Shellar et al. [79] proposed a FL based deep learning model for privacy preservation in multi-institutional collaborations, where the proposed model is trained at all the participating individual institutions without the need to share the individual data among the institutions. In a similar work, Wang et al. [80] proposed a 5G-enabled architecture for auxiliary diagnosis of COVID-19 patients based on FL. In this architecture, multiple institutes can collaborate without compromising on the privacy concerns. The authors in [81] have followed a similar approach for multi institutional collaboration in screening of COVID-19 patients based on chest X-ray images. Li et al. [82] have proposed an FL based framework for analysing the MRI images. The proposed work aims to reduce the time and communication cost involved in transmitting the MRI images from local sites to the central server for storing and training the ML based models.

Even though FL can address the privacy concerns and latency issues in transferring the data from the local devices to the central servers to train the ML models, transparency and tampering of the records still persist. To address these issues, blockchain can be integrated with FL for identifying the patterns from large volumes of medical data while preserving privacy and ensuring security. In this direction, El Rifai et al. [83] have integrated FL and blockchain to train diabetes, kidney, and digestive datasets. FL is used to ensure that

the privacy preservation is accomplished for sensitive data of patients and also to reduce the communication overhead involved in frequently transferring the patients' data from local devices to the central server, whereas blockchain is used to ensure transparency and preventing alterations of the medical records. In a similar work, the authors in [84] proposed a blockchain enabled FL framework to detect the COVID-19 from CT imaging. The data is authenticated by blockchain technology, whereas FL trains the model globally thereby achieving privacy preservation.

### C. Federated Learning enabled Big Data in Smart Transportation

Some of the applications of smart transport systems such as intelligent transportation systems and cooperative autonomous involve sensitive data and large quantity of the devices where the storage, computing, and communication resources must be utilized efficiently. In traditional systems, all the data generated from the vehicles will be transferred to the cloud and the machine learning or other predictive techniques are applied on the data stored centrally in the cloud. However, as the cloud will be making decisions from the global information, large volumes of data from vehicles distributed across several locations have to be acquired and transferred to the cloud. This process needs a high bandwidth and also high delay will be incurred. Also, the decisions should me made based on the local scenarios rather than the global scenarios to take appropriate actions in specific locations. To address the aforementioned issues, FL can be a promising solution as depicted in Fig. 6. FL can be used to train the data from vehicles in a particular area collected and stored at edge devices in near proximity to analyze and extract the patterns from the local area to take immediate actions, preserving the privacy related to sensitive information of the vehicles such as driver details, location of the vehicle, etc. [85].

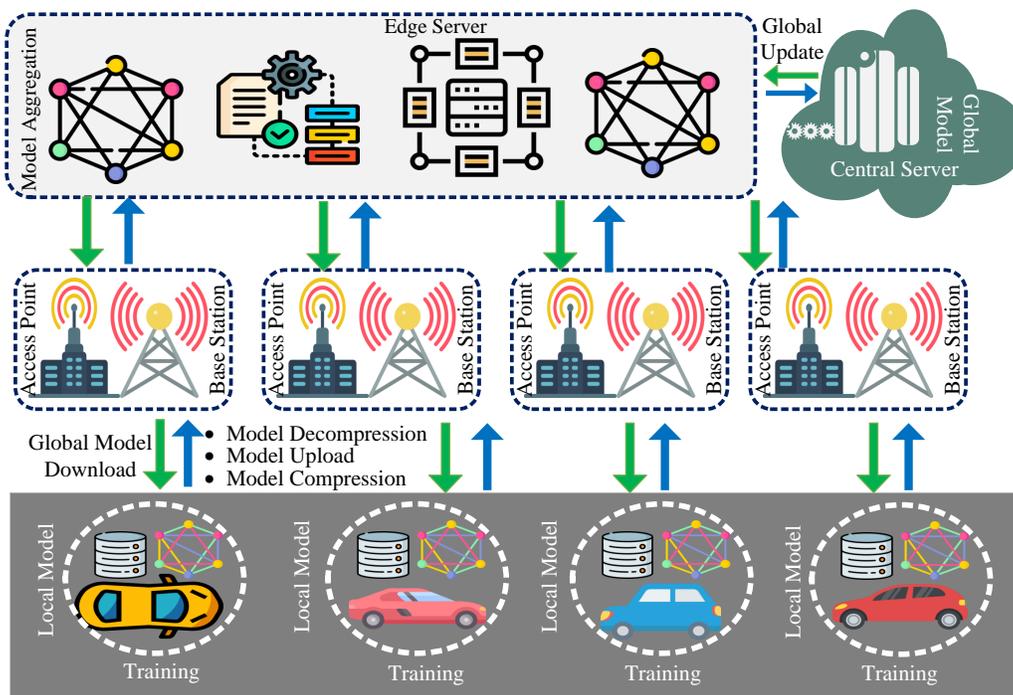


Fig. 6. Federated learning enabled big data in smart transportation.

Road surveillance is an essential application of ITS, where large number of videos generated at fast pace have to be analyzed in real time for the appropriate authorities to take necessary action regarding tasks such as traffic congestion prevention, immediate response to accidents, burglaries on highways, etc. Even though edge computing is promising, new challenges emerge with it like fragmenting the decision models and coordination between the edge nodes distributed in different locations. To address the issues of edge based frameworks when dealing with real time video analytics of ITS, Sada et al. [86] presented a FL based architecture for video analytics which is distributed. The proposed architecture enables distributed and real-time object detection apart from preserving the privacy of the model updates. In a similar work, Xu et al. [87] proposed a FL based traffic monitoring system to obtain the data related to the traffic in roads and other locations using remote sensing data for identifying traffic congestion. FL helps in preserving the privacy of sensitive data related to the vehicles, drivers, etc. In another work, Qi et al. [88] proposed a framework based on FL integrated with blockchain for traffic flow prediction. Even though FL helps in preserving the privacy, it has some serious security issues such as single point of failure of centralized model coordinator. To address this issue, the authors have proposed a consortium blockchain based approach that can provide reliable, secure and decentralized FL without the need for centralized model coordinator. In this work, the miners in the blockchain verify the updates from the models of decentralized vehicles that helps in preventing unreliable model updates. Similarly, Liu et al. [89] proposed FedGRU, a FL-based gated recurrent unit neural network algorithm to predict the flow of traffic for protecting the privacy of large volumes of sensitive data

generated from the vehicles. Mowla et al. [90] proposed an FL based model for detecting on-device jamming attacks in flying ad-hoc networks. The proposed approach helps in reducing the power consumption, communication overhead issues faced by traditional jamming attack detection mechanisms, along with providing on time global updates in flying ah-hoc networks.

In electrical vehicular networks, prediction of the demand of energy is of paramount significance as it ensures accurate charge planning and reliable routing. Saputra et al. [91] proposed an FL based architecture to find the complex patterns in the electrical vehicular networks to improve the accuracy of energy demand prediction. To protect the privacy of sensitive data that will be frequently exchanged between the charging stations and the vehicles, and also to reduce the communication overhead, the authors have used FL that allows the charging stations to share the data without exposing the sensitive data. Similarly, the authors in [92] applied an enhanced FL averaging algorithm that learns regression models and probabilistic neural networks in a privacy-preserving and communication-efficient manner to address the predictive uncertainty of traditional FL based models.

#### D. Federated Learning enabled Big Data in Smart Grid

The electric grid concept is aging and the current electric grid infrastructure is finding it difficult to handle the things it is not designed for. Using cutting edge technologies like IoT, ML, communication networks, equipment in electric grids will ensure that the electricity is delivered more efficiently and reliably. The modernized electric grids, termed as smart grids, will result in reducing the frequent and long power outages, faster restoration of services when the outages occur, reduce the impacts of the storms on the electricity supply, etc.

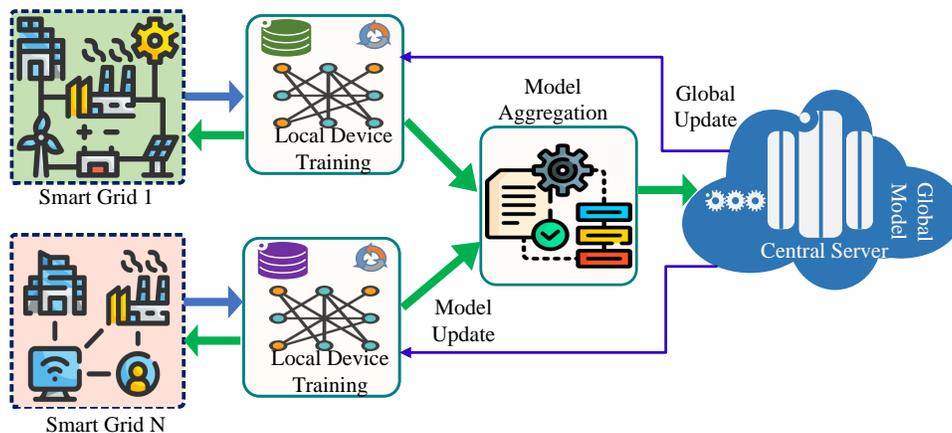


Fig. 7. Federated learning enabled big data in smart grid.

Smart grids generate large quantity of data at a very rapid pace. ML algorithms can be used on this data to analyze the patterns and predict the electricity demand, power outages, stability of the smart grid, electricity theft, etc [93]. However, the traditional approach of transferring the data from the smart grid to the central cloud and then applying ML algorithms involves several challenges such as increased latency, exposing of sensitive and private data of the customers/grid to the potential hackers, etc [94]. FL, through its inherent characteristic of training the global ML model in the local devices and sending only the model parameters to the central ML model for training, has a great potential to address these issues of the smart grid as depicted in Fig. 7. The recent literature related to application of smart FL for smart grids is discussed in rest of the subsection.

Deep learning models can expose privacy sensitive data such as the data related to the appliances collected by a smart meter, that may reveal the consumer's behavior at home based on the appliances used by them, that can raise serious security concerns. Also, the ML/deep learning models need large quantity of data to get trained. To address these issues, Taik et al. [95] proposed a model based on FL and edge computing for household load forecasting that addresses the volume and diversity requirements of the deep learning models and also preserves the privacy of consumer data. In a similar work, Wang et al. [52] proposed a FL based method to identify the electricity characteristics of the distributed consumers that can preserve the privacy of the retailers.

The future energy demand of a smart home can be predicted by training a deep learning model on the energy consumption data of a consumer. The data collection from the clients who are distributed for centralized model training incurs huge costs with respect to communication resources. To address this issue, Tun et al. [96] proposed a FL based model, in which every client has to upload only the model updates generated by training the FL model on its local data and these updates are then aggregated into a global model at the centralized server.

Forecasting the load of a power system in short term play an important role in the efficient, stable management and dispatch of the electricity. The traditional model of centralized training of the ML based models for forecasting the load of

the power system has issues regarding the privacy preservation and also the latency incurred due to the large volumes of data generated from distributed consumers. To address these issues for electric load prediction, Li et al. [97] proposed an FL based approach for short-term load forecasting of the grid. In a similar work, Savi et al. [98] proposed a framework based on edge-computing and FL that use LSTM models for short-term consumption of energy of the consumers.

#### E. Federated Learning enabled Big Data for Online Recommender Systems

Recommender systems are seen commonly these days in several online applications such as amazon, youtube, netflix, etc. Recommender systems suggest items that are relevant to the users, such as products to buy, text to read, movies to watch, songs to listen, etc. Typically, the recommender systems have to process large volumes of data related to the activities of the users to filter, rank and suggest/recommend the items based on their interests [99]. Users browse large volume of items on a daily basis. Analyzing this big data in a traditional approach, where the data is collected from the users and stored centrally for training an ML based model will incur heavy communication and privacy costs. FL can play a vital role in addressing the aforementioned issues related to big data from recommender systems. Some of the recent state-of-the-art on the application of FL for recommender systems that have to deal with the big data are discussed below.

Amir et al. [100] proposed an FL-based model to improve the personalized recommendations to the users in a privacy preserving manner. The proposed system ensures that the ratings of the users on the items remain in their local device, and only the preference patterns are shared to the global model that reduces communication overhead. In a similar work, Tan et al. [101] proposed an FL based approach that trains the data from several clients, while at the same time, preserving the privacy of each client. In another interesting work, the authors in [102] used a FL based model to preserve privacy of the user's preferences in a recommender system. The users have tested their proposed approach on Alibaba's e-commerce recommendation and evaluated their model on

Taobao user data collected over 30 days. The results proved that the proposed system's scalability, reduced communication, storage and computational overhead, while preserving the privacy of the user's data simultaneously. Muhammed et al. [103] presented a novel FL approach in which, a diverse set of sample data from different clients is trained on the federated model and the model updates are propagated to the other clients using an active aggregation technique. In this way, the users can benefit from the proposed approach with reduced communication costs and also the models which are more accurate and can be consumed by the users at even very early stages of training.

#### *F. Federated Learning enabled Big Data for 5G and 6G*

FL trains machine learning models with devices or relevant data centers and at the same time maintains these training datasets locally in the system avoiding the possibility of sharing raw data. This procedure of FL implementation is usually divided into three phases. The first phase is the initialization phase wherein the central aggregator represents the global model which is pre-trained on a public dataset to each of the devices. Then the devices use the 5G network to train and further improvise the global model on the dataset for each iteration. The local model updates get aggregated by the central aggregator in the aggregation phase followed by the update phase wherein a new global model gets generated to be used in the next iterations. This process is iterated until the optimum accuracy and convergence level is achieved by the global model. The model proves to be much secured with lesser risk of privacy leaks due the decoupling of the training of the model from access to raw data used during the training phase. It is due to the advent of fifth-generation (5G) cellular networks that the number of wireless devices will increase stupendously reaching the scale of trillions by the next decade. The advantages of 5G includes enhanced coverage, transmission bandwidth along with reduction in communication latency in versatile applications of smart homes, autonomous vehicles, UAVs and various others. The average 5G speed will soon be much higher than the average mobile connectivity and this has resulted in the announcement of over 81 billion dollar bidding for 280MHz licenses of aforementioned frequency. In order to use this valuable spectrum resource, the third generation partnership projects (3GPP) have been trying to standardize and enable 5G new radio services in unlicensed spectrum being part of LTE which is a Licensed Assisted Access initiative (LAA). FL acts as a hybrid technique involving centralized and decentralized learning. The model training is done using the decentralized technique and the data gets stored at its source. The raw data is stopped from being transferred to prevent communication overhead thereby optimizing performance and security [72], [104], [105].

#### *G. Federated Learning enabled Big Data for Industry 4.0/5.0*

FL has evolved as an extremely successful solution for the development of cost effective and secured IIoT applications. The implementation of FL helps in attracting large dataset and computation resources from various IIoT devices for the

training of the AI models. This improves the quality of IIoT training data which may not be possible to be achieved using the traditional AI approaches. The integration of FL with IIoT have various advantages. Firstly, as already discussed, in an FL system the local updates are only accessed by the central server but the local data are kept safe in the local devices. Hence as per the General Data Protection Regulation (GDPR), the FL successfully develops secured, sustainable and safe IIoT systems. Also, since offloading of stupendous data volumes to the server is avoided in FL, the communication gets significantly reduced thereby saving network spectrum resources in the IIoT networks. This motivates the use of FL in applications for mobile devices, manufacturing robots, industrial automation and various others [17], [106].

#### *H. Federated Learning enabled Big Data for Social Media*

These days most of the people are spending great amount of their time on their mobile phones, especially, social media. In several social media apps such as Facebook, Instagram, Twitter, etc., the users will be posting photos, videos, tweeting their status, sharing their opinions, sharing other posts, etc., that results in generation of heterogeneous data at a very fast rate and of enormous proportions [107]. Traditionally, the data from the mobile users used to be transferred to the central cloud to train the ML based models for gaining useful insights from these social media posts. However, this may expose the sensitive information of the users such as their preferences, location, etc. Also, large quantity of communication resources are required to share the data to the central cloud, that would in turn introduce latency. To address these issues, FL can be utilized to train ML models on the user's mobile devices itself, that would result in privacy preservation of the social media users, optimize the resources, and improve the real-time analytics of the big data related to social media [108].

#### *I. Federated Learning enabled Big Data for Other Applications*

FL can be used for several other applications that are based on big data such as pollution control, defense etc. Liu et al. [72] proposed a novel FL based framework for monitoring and predicting the quality of the air. The proposed framework predicts the air quality index scale distribution by using a Dense-MobileNet model that learns the haze features from the images taken by Unmanned Aerial Vehicles (UAV) in the air. The traditional method of transferring the large volumes of the data generated by the UAVs incurs heavy communication overheads. Also, the generated data may include sensitive information that should not be exposed. The proposed FL enabled architecture ensures several organizations to learn collaboratively through a global model to monitor air quality index in a privacy preserving manner. The authors proposed a long short-term memory model based on graph convolutional neural networks for ground sensing systems to achieve real-time, accurate and future air quality index inference. FL enabled technologies can be effectively used by UAV's in several big data applications such as predicting optimal routing path, forecasting the mobility patterns on the ground users,

monitoring in crisis/disaster situations, defense etc [109], [110]. FL enabled big data frameworks can also be used in several applications such as online shopping, insurance, credit card divisions etc., where the privacy of the personal information of the users has to be preserved.

### J. Summary

From the discussion above, it can be summarized that FL has a tremendous potential in addressing several challenges faced by big data applications such as privacy preservation, management of resources to handle large volume of data, communication cost and latency issues involved in transferring large volumes of data at a rapid pace to the central server for training the ML algorithms, real time analytics, customized decisions based on geographical locations, and heterogeneity of the data. However, several challenges such as poisoning of ML models in local devices, presence of malicious terminals that may result in wrong training of FL models, and excess energy requirement at local devices have to be addressed to realize the full potential of FL in big data applications. Table III summarizes the benefits and challenges of FL in big data applications.

## V. FEDERATED LEARNING BIG DATA PROJECTS

This section presents the key FL projects and platforms which are widely used for Big Data developments.

### A. Flower: A Friendly Federated Learning Research Framework

Flower is an open-source platform-independent FL framework<sup>1</sup>. Flower framework is specifically designed to provide scalability, which is appropriate to handle big data applications. Flower framework has been used for many real-world applications, which can also support more than 10,000 clients. Moreover, Flower framework supports a wide range of devices, including Android mobiles, iOS devices, Raspberry Pi and Nvidia Jetson.

Key features of Flower framework are as follows [111].

- **Support for heterogeneous clients:** Flower can offer the same workload on clients running on different operating systems using different languages.
- **Provide High Scalability:** by allowing the workload to scale across thousands of machines.
- **Separation of Client-side and Server-side definition** which allows independent control over local client-side and global server-side computations.
- **Support of ML framework-agnostic libraries:** allows the users to use their favorite and experienced ML frameworks
- **Language-agnostic support:** allows users to implement FL clients in different languages and emerging embedded device platforms
- **Support for baselines:** allows fast comparison of the new FL algorithms with existing and well-known algorithms.

<sup>1</sup><https://flower.dev/>

### B. Pysyft

PySyft is one of the most popular open-source deep learning frameworks which offer privacy and support FL<sup>2</sup>. It is a Python library to perform secure and private deep learning methods [112]. Moreover, PySyft can be integrated with other deep learning frameworks such as PyTorch, Keras, or TensorFlow.

PySyft framework is leveraging three main techniques, i.e., Secured Multi-Party Computations (sMPC), differential privacy, and FL. The use of sMPC in PySyft helps to protect the privacy of input data. sMPC can be used to perform computations over private inputs data. Differential privacy can reduce the statistical privacy leakages in large dataset. In addition, PySyft also supports Homomorphic Encryption (HE) to offer extra privacy.

### C. Fate: An Industrial Grade Federated Learning Framework

Fate is another open-source FL framework developed by Webank's AI department<sup>3</sup>. Fate platform primarily supports big data collaboration according to the regulations. It was enabled by integrating multiple secure computation protocols in Fate framework. Fate platform uses several features such as a flexible scheduling system, a modular, scalable modeling pipeline, and clear visual interfaces to keep the scalability, user-friendliness, and improved operational performance.

Fate platform support different service modules such as

- **FederatedML:** supports common FL/ machine learning algorithms and tools such as DataIO, Intersect, Federated Sampling, Feature Scale, Hetero Feature Binning, OneHot Encoder, Hetero Feature Selection, Hetero LR, Hetero Poisson, Homo LR, Homo NN, and Hetero Secure Boosting.
- **FATE-Serving:** support high-performance FL algorithms.
- **FATE-Flow:** is an end-to-end pipeline platform to perform highly flexible and high-performance FL tasks.
- **FATEBoard:** is a tool to visualize the FL models. It is useful to explore, improve and debug the FL models efficiently.
- **FATE Network:** is designed to enable secure and efficient communications between FL parties.
- **KubeFATE:** manages the workload of FL tasks to achieve high scalability, high efficiency, and cost reduction.

### D. FedML: Federated Learning Library

FedML is an open research library and benchmark which supports the deployment of evaluation of novel FL algorithms [113]. FedML supports standalone simulation, distributed training, and mobile on-device training like many other FL platforms. FedML supports worker/client oriented programming to offer an easy user experience. Moreover, it enables end-to-end design pattern toolkits to compare and evaluate novel custom FL algorithms with existing known algorithms.

<sup>2</sup><https://blog.openmined.org/tag/pysyft/>

<sup>3</sup><https://fate.fedai.org/>

TABLE III  
BENEFITS AND CHALLENGES OF FL IN BIG DATA APPLICATIONS.

Application	Existing Challenges Faced	Benefits of using FL
Smart City	1) Privacy Preservation of sensitive data of citizens, vehicles 2) Resource Management 3) Communication cost in transferring large volumes of data 4) Providing specific solutions relevant to varied city demographics and needs is difficult to disseminate during emergency periods in real time	1) Private data is not shared with the central server; instead model updates are shared 2) Since the FL model is trained on local devices, burden of resource management in the central cloud is reduced 3) As the data is trained locally by a federated model and only model updates are shared with the central ML model, cost of communication is significantly reduced 4) Using FL, suitable solutions can be provided to the diverse needs in different areas of the city based on relevant data
Smart Health-care	1) Privacy preservation of the sensitive data of patients 2) Real time analytics during pandemics 3) Sharing large volumes of medical data from the sensors, wearables to the central cloud 4) Generating diverse data from different medical institutions for effective training of ML models	1) By using FL, the ML model can be executed in local devices, hence privacy of the data of patients can be preserved 2) FL can provide analytics in near real time during pandemics as the local device need not wait for the arrival of data from other devices to perform the ML 3) The FL models can reduce the communication overhead as only local updates are shared to the central server 4) Collaboration between multiple hospitals without exposing the private and sensitive data of patients is possible through FL, where global model can be executed at individual hospitals without actually exposing the sensitive data of the patients
Smart Transportation	1) Preserving privacy of sensitive information of drivers, vehicles 2) Latency and communication cost involved in transferring the large volumes of data from the vehicles to the cloud 3) Customized decisions based on the traffic information in a particular area	1) FL can ensure privacy as only model parameters will be shared with the central server. 2) As FL is executed in local devices/vehicles latency and communication cost issues will be resolved 3) FL can enable personalized decisions based on the traffic information of that specific area as FL is trained based on the local data
Smart Grid	1) Privacy reservation of the consumers and appliances 2) Increased latency and communication costs in transferring large volumes of the grid data to the central cloud 3) Real time analytics on the smart grid data	1) Since only model parameters are shared by FL, privacy of sensitive information of the customers and the appliances remain protected 2) The latency and communication costs involved in transferring large quantity of the consumer data are reduced as the FL models get executed in local devices at consumer location 3) Since FL models are executed on local edge devices, it enables efficient and timely decisions without waiting for data from other locations
Online Recommender Systems	1) Privacy preservation of the users 2) Communication costs involved in transferring large volumes of data to the central cloud	1) The sensitive information of the online users will not be exposed as FL will be executed on local devices and only model updates are shared to the central server 2) As FL is executed in local devices, and only model parameters are shared to the central server, the communication costs can be reduced
Social Media	1) Privacy preservation of the sensitive data of the social media users 2) Communication costs involved in sharing large volumes from social media at rapid pace	1) The sensitive data of social media users namely their patterns in posts, their moods, and location remain unexposed to the central server as model parameters alone are shared by the FL 2) Since FL is executed in local devices, and only model parameters are shared to the central server reduction in communication costs can be observed.

### E. TensorFlow

TensorFlow<sup>4</sup> is also widely used open-source machine learning platform. The Google Brain team originally developed it for their internal use. TensorFlow supports a wide range of flexible tools which can be used to deploy ML-powered applications and build customized ML models.

TensorFlow Federated (TFF)<sup>5</sup> is extension of TensorFlow platform. TFF is specifically designed to support decentralized data processing and FL. TFF contains a rich set of tools that users can use to run FL algorithms on their models and data. Moreover, TFF supports the experimentation of new FL algorithms.

There are two layers in the TFF platform.

- **FL Layer:** It provides the interfaces to interconnect existing Keras or non-Keras machine learning models with the TFF platform. This layer supports the users in conducting basic FL tasks such as FL training and FL evaluation. Here, users do not need to be fully aware of the mechanism or structure of the FL algorithm to run the tests.
- **Federated Core (FC) Layer:** This layer support customizes user FL algorithms. It offers low-level interfaces which can be used to interconnect TensorFlow with distributed communication operators and perform custom FL algorithms.

<sup>4</sup><https://www.tensorflow.org/>

<sup>5</sup><https://www.tensorflow.org/federated>

### F. LEAF: A Benchmark for Federated Settings

LEAF is a popular modular benchmark framework for FL settings. It supports the benchmark of FL's different applications, including multi-task learning, meta-learning, and on-device learning<sup>6</sup>.

LEAF framework includes three main components [114].

- **A suite of open-source datasets:** This is responsible of preprocess the data and covert to standardized format. LEAF support different well-know data sets such as Federated Extended MNIST (FEMNIST) [115], Sentiment140 [116], Shakespeare [117], CelebA [118], Reddit and a Synthetic dataset [119]
- **An array of statistical and systems metrics:** This contains the different evaluation metrics which can be used to assess the behaviour of learning solutions in federated scenarios.
- **A set of reference implementations:** This contains a set of reference implementations such as Mocha [120], FedAvg [121] and minibatch SGD, which support FL.

Thus, LEAF allows researchers in FL to benchmark their novel and customized FL solutions under realistic settings. The comparison the popular FL platforms is summarized in Table IV.

<sup>6</sup><https://leaf.cmu.edu/>

TABLE IV  
COMPARISON OF EXISTING FL PLATFORMS [111] [113] [112] [114]

		FL Platforms					
		Flower	Pysyft	Fate	FedML	TFF	LEAF
Algorithms	SplitNN	Yes	No	No	yes	No	No
	FedNAS	Yes	No	No	Yes	No	No
	FedAvg	Yes	Yes	Yes	Yes	Yes	Yes
	VFL	Yes	No	Yes	Yes	No	No
	decentralized FL	Yes	No	No	Yes	No	No
Computing Platform	On-Device	Yes	No	No	yes	No	No
	Standalone	Yes	Yes	Yes	Yes	Yes	Yes
	Distributed	Yes	Yes	Yes	yes	Yes	No
Benchmark	vertical FL	Yes	No	Yes	Yes	No	No
	Shallow NN	Yes	Yes	Yes	Yes	Yes	Yes
	Model DNN	Yes	No	No	Yes	No	No
	Linear models	Yes	Yes	Yes	Yes	Yes	Yes
API Design	Customizable typologies and message	Yes	Yes	No	Yes	No	No
	Message flow flexibility	Yes	No	No	yes	No	No

## VI. RESEARCH CHALLENGES AND FUTURE DIRECTIONS

As already discussed, FL is a machine learning based technique wherein various clients collaboratively perform training of a model under the control of a centralized server but the training data remains decentralized. Machine learning algorithms, particularly deep neural networks get trained using multiple datasets being part of local edge nodes. The raw data in a traditional system gets aggregated to a centralized data centre for the purpose of training. But in case of FL the raw data is left distributed on the client devices being further trained as a shared model on the server through aggregation of local computed updates. There exists several research challenges associated with FL in big data applications. These challenges can be broadly dichotomized into two categories namely the training related challenges and security related challenges. The data training related challenges include communication overhead while performing multiple training iterations. It also includes issues relevant to heterogeneity of the devices used in learning and heterogeneity of the data used in training. On the other hand, security challenges include threats to privacy and security caused by adversaries starting from malicious clients in the local device to malicious users who has black-box access to a model. In case of FL the private data remains inside the device instead of leaving and thus makes it easier for an unauthorized user to learn about the existence of data point used for the training in the local models. Security attacks are often introduced due to the manifestation of targeted or non-targeted malicious attacks in the learning process. In case of the targeted attacks, the adversary manipulates the labels of specific tasks. In case of non-targeted attacks, the attacker primarily intends to compromise the accuracy of the global model [122]. Some of the core challenges in the implementation of FL in big data applications are discussed in the next sections as illustrated in Fig. 8.

### A. Bottleneck in Communication Efficiency

Communication emerges as a bottleneck in the development of FL based systems. The primary reason for this is the inclusion of massive number of devices and communications in the network which could be slower than the local computation of other magnitude orders. FL basically sends small messages or model iteratively being part of the distributed training process instead of transporting the entire dataset over the network. The following strategies act as potential concepts to achieve communication efficient learning methods [122]–[124].

#### Possible Solutions and Future Direction of Research:

The first strategy is to use Local updating methods that allows variable number of local updates to be implemented on each machine in parallel during each communication round. This would help lessen the number of communication rounds. Secondly model compression mechanisms namely sparsification, subsampling and quantizations also helps to reduce the message sizes that get communicated at each update round. Thirdly, in case of FL the server connects to all remote devices in the setting. Hence decentralized topologies act as an alternative solution during the instances of bottlenecks while operating with low bandwidth and high latency networks.

Federated Optimization: There exists various questions relevant to federated optimization in dealing with IID data. There are research gaps between the upper and lower bounds of optimization in FL settings when using intermittent communication graphs which capture the local SGD approaches. But the convergence rates of these approaches do not tally with the relevant lower bounds. Although, majority of the schemes are able to successfully achieve the asymptotically dominant statistical term, but all of them fail to achieve the convergence rate of the accelerated mini-batch SGD. Hence there lies scope of research that would enable federated averaging algorithms to eliminate the gap between the optimization bounds [122]–[124].

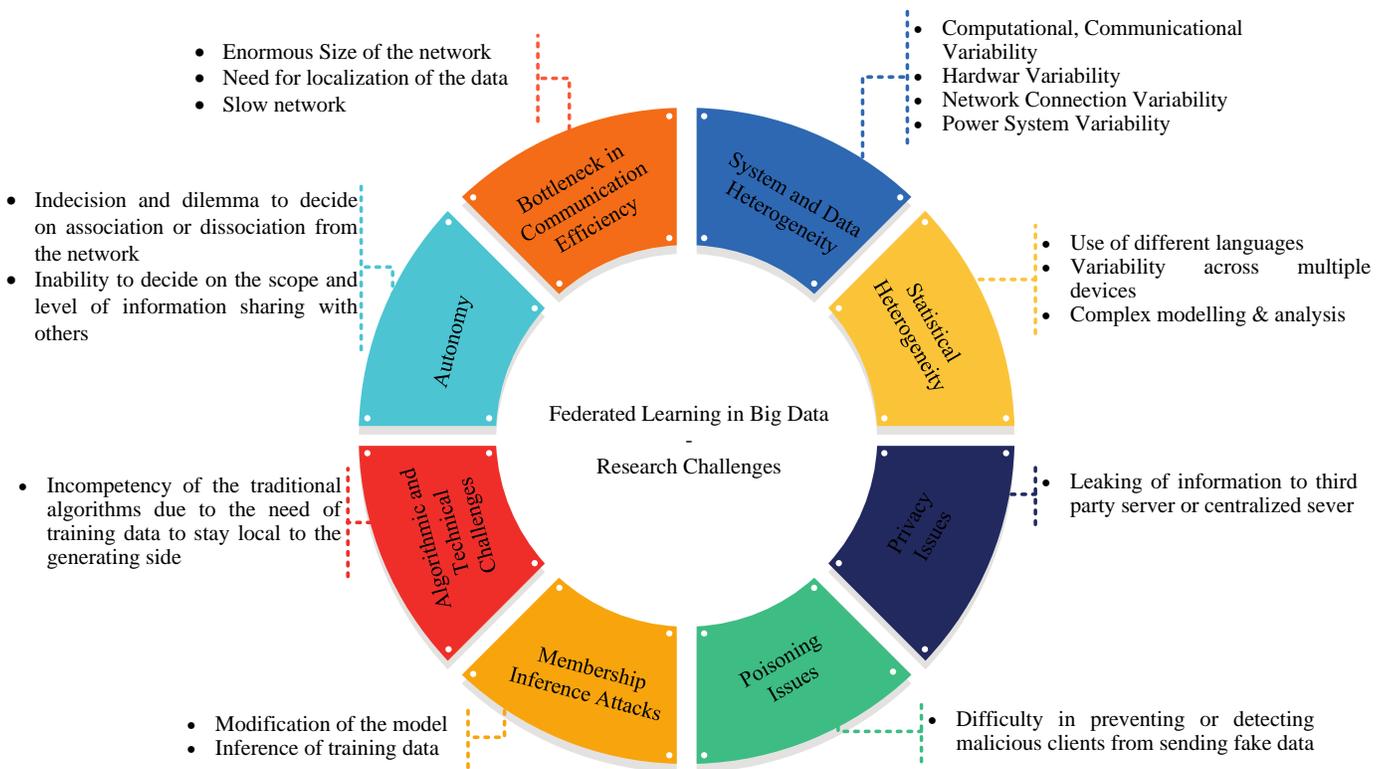


Fig. 8. Federated learning in big data-research challenges.

### B. System and Data Heterogeneity

The capability to store, compute and communicate using the devices being part of federated network vary significantly from one another. The reason behind such differences are related to hardware, network connection and power supply. Also only a small part of the data should be active at any point in time. Here each device has the possibility of being unreliable because it is quite common for edge computing devices from the network due to connectivity or energy based limitations. The non - identically distributed data collected from the various devices have significant effect on the performance of FL systems. The issues pertinent to system heterogeneity can be achieved implementing the following strategies. Firstly, the use of asynchronous communication to parallelize the functioning of iterative optimization algorithms. This also helps to eliminate stragglers in heterogeneous environment. Active sampling of devices at each round while aggregating device updates within pre-defined window is a plausible solution to deal with heterogeneity. Ignoring of fault tolerance and inclusion of algorithmic redundancy as part of code computation also helps to achieve algorithmic redundancy [12], [122]–[124].

#### Possible Solution and Future Direction of Research:

To avoid system heterogeneity, asynchronous communication technique is used which helps to run the iterative optimization algorithms in parallel. These act as a highly potential approach to eliminate possibilities of stragglers in the heterogeneous

environment. The second approach involves selection of participating devices at each round actively to ensure maximum aggregation of updates within a pre-defined window wherein only a small subset of devices participate at the end of each round of training. The third plausible approach is to avoid device failures leading to bias in the device sampling scheme when the failed devices have some specific characteristics of data. Also algorithmic redundancy could be introduced as part of coded computation technique to achieve fault tolerance [12], [122]–[124].

### C. Statistical Heterogeneity

As already mentioned, the various devices in the FL systems, collect data in a non-identically distributed manner across the network as per the varied types of uses by the end users. The data point counts thus may vary significantly due to the presence of underlying structure that captures the connectivity among the devices and their relevant distributions. The data generating paradigm increases the chances of having stragglers that add complexity in the modeling, analysis and evaluation in the system. The issues arise while training the federated models from the data that are non identically distributed both in terms of data modeling and analysis of the convergence behaviour relevant to associated training procedures [12], [57]–[59], [122]–[124].

#### Possible Solution and Future Direction of Research:

Heterogeneous Diagnostics: There exists studies focusing on quantification of statistical heterogeneity using different metrics which are mostly calculated during the training phase.

This motivates the need for developing simplistic diagnostic techniques for the quantification of statistical and system heterogeneity even before the training phase. Also there lies scope of research to identify ways to improve the convergence techniques involved in federated optimization methodologies [12], [57]–[59], [122]–[124].

#### D. Privacy Issues

The privacy concerns persuades users to keep the raw data in each devices localised in the federated settings. But sharing of various other information included in the training process has the potential to disclose sensitive information to the third party of the centralized server. The secure multiparty computation (SMC) or differential privacy techniques contributes towards achieving privacy in FL. But there are associated challenges of deteriorated performance and efficiency in the model. The management of balancing all of these security trade-offs act as a significant challenge in achieving privacy in FL systems. In case of FL, the model updates alone namely the gradient information is shared which ensures data protection on each device. But in spite of such attempt, there are open possibilities of sensitive information being shared with third parties or the centralized server while processing the updates of the communication model. Most of the development in this regard emphasizes on the use of cryptography techniques or implementation of differential privacy. But these techniques also have their individual challenges which motivates further development of innovative methodologies to achieve data privacy. As an example, some studies include strategy in which all possible information about counter party's data is acquired abiding to all the necessary protocols. There also exist techniques which allow involvement of malicious parties where they can tamper the system by feeding falsified input [12], [122]–[124].

#### Possible Solutions and Future Direction of Research:

The concept of privacy lies normally looked upon from the local or global level considering all devices in the network. The future direction of research lies in defining the concept of privacy at a much granular level as the various privacy related constraints differ at various devices or even the data points in a single device. The objective lies in developing privacy techniques capable of handling mixed privacy restrictions [12], [122]–[124].

#### E. Security Issues

*Poisoning*:: There exists possibilities of two types of poisoning namely Data Poisoning and Model Poisoning in FL systems. Data poisoning in FL based training process occurs when multiple clients participate in the contribution of their on-device training data. In such circumstances, it becomes immensely difficult to stop or detect malicious clients from sending falsified fake data. These fake data poison the training process and the model also gets poisoned. In case of Model poisoning, the malicious client tampers the gradient or parameters which further modifies the received model before being sent to the centralized server for integration of information.

The entire global model gets poisoned during this aggregation process where invalid gradients get included. With the increase in model dimensionality the possibility of poisoning and backdoor attacks further increases. The plausible solution in this regard would be to share only prediction results or information which are less sensitive leading to achievement of a more robust and protected method ensuring optimum level of privacy [12], [57]–[59], [122]–[124].

#### Membership Inference Attacks:

The raw data is usually stored in the local device yet there exists chances of inferring the training data in an FL based system. As an instance, the information regarding the training data can be inferred during the process of updating the model in its learning process. The defence technique actually focuses on achieving differential privacy in this regard. The most common approach adopted are focused on ensuring secured computation, differential privacy and development of a trusted execution environment. In the process of achieving secure computation, two main techniques are adopted namely - Secure Multiparty Computation (SMC) and Homomorphic Encryption. In case of SMC, it is agreed to perform the inputs by two or more number of parties and the output is revealed to only a subset of the participants. On the contrary, in case of homomorphic encryption, the computations are implemented on encrypted inputs before they getting them decrypted. In the process of implementing differential privacy scheme, the user's contribution is masked. Noise is added to the clipped model parameters before integrating the model although it leads to compromising of the model accuracy. The trusted execution environment presents a secured platform to run the FL process in low computation overhead which is presently suitable only for CPU devices [12], [57]–[59], [122]–[124].

*Model Inversion Attacks*: In this type of attack the training data gets compromised. The potential attacker gets access to the target labels and queries the final trained model. The returned classification scores get exploited in order to reconstruct the remaining data.

*Membership Inference Attacks*: Here the attackers try to find out if some data has been part of the training. Shadow models are created exploiting the returned classification scores with similar classification boundaries as the original model which was subjected to attack

*Model Encoding Attacks*: In this case, the attacker having white-box access attempts to find the training data which had been memorised by the model weights. On the contrary, in case of a black-box situation the original training model gets over-fitted for getting unauthorized access to parts of the training labels.

*Model Stealing Attacks*: In case of model stealing attack, the attacker intends to steal the entire model. When the model is sent to the participants for the purpose of training, a duplicate model mimicking the original decision boundaries get created. Thus the attackers escape payment of usage fees to the ML experts of the original model or while selling the model to the third parties [12], [57], [122]–[124].

**Possible Solutions and Future Direction of Research:** One

of the possible mitigation steps to security concerns is the implementation of differential privacy schemes. In this case, the user's contribution before aggregating, is masked by the addition of noise to the clipped model parameters. Although the accuracy of the model gets negatively affected due to the noise, but it acts as a mathematical guarantee of achieving algorithmic output even when a particular data gets used for the training process [12], [57], [122]–[124].

The use of Secure Multi-Party Computation (SMPC) function allows multiple participants to derive a process mutually similar to the ML training procedure. In such case, only the outcome of the function gets disclosed to the participating members but not the relevant training information. The gradients and the parameters gets computed, updated and encrypted through a decentralized process. The custody of each data item are spit into shares that are sustained by the relevant participating entities. The SMPC although helps in achieving privacy during training period but remains vulnerable during the testing phase [12], [57], [122]–[124].

Homomorphic Encryption also acts as a mitigation technique wherein a complex cryptographic protocol is implemented that ensures mathematical derivation of the encrypted data and the generated outcome also remains encrypted. HE has the potential to ensure security in both training and testing phase. The possible future direction of research lies in making this intensive computation technique economical to be fit for application in real world scenarios [12], [57], [122]–[124].

#### *F. Trust Issues*

In FL based frameworks, the ML model owners send their respective models to the data holders to implement training. Apparently this may seem to be secure but there exists trust issues. To highlight some of the examples, the ML model owners could alter their model and get information on their training data. Apart from this, in a FL setup although the data is not shared in the original format, it has risks of being reconstructed in case there exists lag in the protection of architecture and related parameters. There is also a possibility of FL exposing intermediate results namely the parameter updates from the optimization algorithms like stochastic gradient descent (SGD). The transmission of these gradients could also lead to leaking of private information when it gets exposed with image pixel data structure. Hence protection of parameters still remains to be an area of concern while striking a balance between privacy-security and achievement of system performance [12], [57], [122]–[124].

**Possible Solutions and Future Direction of Research:** The potential solution to the aforementioned trust issues could be the implementation of secure aggregator. Studies could be done in the development of automated middleware procedure or program which would aggregate all the trained models and send the updates to the ML model owners acting as a robust model against various trust compromising attacks [12], [57], [122]–[124].

#### *G. Algorithmic and Technical Challenges*

The traditional approaches do not work well in the FL paradigm which triggers the proposition of new algorithms

such as Federated Averaging (FedAvg) and Federated Stochastic Variance Reduced Gradient (FSVRG) and the Co-Operative machine learning model. In FedAvg the entire training process is controlled by the centralised server which encompasses the shared global model enabling overall communication. Stochastic Gradient Decent Method is used for optimization to be done on the clients locally. In case of the FSVRG technique, the full gradient computation is performed centrally and the other distribution updates are done at each client by performing random permutation on local data iteratively. The FedAvg and FSVRG approach both focus on updating the model undergoing a synchronized approach for model updation. On the contrary, the Co-operative machine learning model takes an asynchronous approach which integrates the received client model with the global model. Although all the three techniques mentioned work well in case of IID and non - IID but their potential in case of unevenly distributed data in real time environment is still unsure [12], [57], [122]–[124].

#### **Possible Solution and Direction of Future Research**

**Development of Innovative Models of Asynchrony:** The two predominant communication schemes explored in distributed optimization are the bulk synchronous and asynchronous techniques. On the contrary, in case of federated networks, the devices often remain unassigned to the tasks in process hence most devices remain inactive during any of the iterations. Hence as part of future research there lies opportunity to develop device-centric communication models. These models work beyond the synchronous or asynchronous training schemes and individually each device decides on the time to interact with the server rather than being subjected to the workload [12], [57], [122]–[124].

**Threshold level Communication Schemes:** The knowledge on optimum level or threshold level of communication required in FL is still unexplored. There lies opportunities of research to acquire detailed theoretical and empirical idea relevant to one-shot and few-shot communication schemes in the implementations of massive statistically heterogeneous networks [12], [57], [122]–[124].

#### *H. Autonomy*

Generally in a FL environment, all communication devices tend to be autonomous. In this regard, Association Autonomy deals with the autonomy and willingness of the devices to join or drop out of the network and also provide independence to participate in one or more networks. In case of Communication Autonomy, the device has the freedom to decide on communicating with other participant devices and also the scale of data participation. Although FL is quite robust when communicating with any device while joining or leaving a system, but the devices may also choose to make agreements to seamlessly enter or exit from an FL system. The Google FL system and blockchain are examples that contribute to the same purpose [12], [57], [122]–[124].

**Possible Solutions and Directions of Future Research:** **Adaptation of Centralized Training Workflows:** The inclusion of centralized training workflows such hyper-parameter tuning, neural architectures, debugging and interpretability in the FL

setting acts as a bottleneck in extensive adoption of FL in real time setting. Hence this acts as a potential area of future research [12], [57], [122]–[124].

## VII. CONCLUSIONS

Due to the unprecedented growth of IoT and generated data, there are many big data services and applications. However, conventional approaches using traditional AI/ML techniques for big data face critical issues, such as, data privacy, data variety, communication efficiency, and scalability. In this context, FL is an AI breakthrough in the implementation of big data services and applications. As a result, this paper sets to provide a comprehensive survey on the use of FL for various big data services and applications. We have started by providing the fundamentals of big data and FL, and the motivations of the integration of FL into big data. Then, we have reviewed that FL is a promising AI technique to overcome the challenges in big data services and applications. Finally, from the extensive review, we have highlighted several key challenges of this topic and discussed a number of interesting directions.

## REFERENCES

- [1] A. Oussous, F.-Z. Benjelloun, A. A. Lahcen, and S. Belfkih, "Big data technologies: A survey," *Journal of King Saud University-Computer and Information Sciences*, vol. 30, no. 4, pp. 431–448, 2018.
- [2] M. Favaretto, E. De Clercq, C. O. Schneble, and B. S. Elger, "What is your definition of big data? researchers' understanding of the phenomenon of the decade," *PloS one*, vol. 15, no. 2, p. e0228987, 2020.
- [3] M. Mostert, A. L. Bredenoord, M. C. Biesaat, and J. J. Van Delden, "Big data in medical research and eu data protection law: challenges to the consent or anonymise approach," *European Journal of Human Genetics*, vol. 24, no. 7, pp. 956–960, 2016.
- [4] A. Mohamed, M. K. Najafabadi, Y. B. Wah, E. A. K. Zaman, and R. Maskat, "The state of the art and taxonomy of big data analytics: view from new big data framework," *Artificial Intelligence Review*, vol. 53, no. 2, pp. 989–1037, 2020.
- [5] B. Ristevski and M. Chen, "Big data analytics in medicine and healthcare," *Journal of integrative bioinformatics*, vol. 15, no. 3, 2018.
- [6] A. L'heureux, K. Grolinger, H. F. Elyamany, and M. A. Capretz, "Machine learning with big data: Challenges and approaches," *Ieee Access*, vol. 5, pp. 7776–7797, 2017.
- [7] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: A new perspective," *Neurocomputing*, vol. 300, pp. 70–79, 2018.
- [8] M. I. Baig, L. Shuib, and E. Yadegaridehkordi, "Big data adoption: State of the art and research challenges," *Information Processing & Management*, vol. 56, no. 6, p. 102095, 2019.
- [9] F. M. Awaysheh, M. Alazab, S. Garg, D. Niyato, and C. Verikoukis, "Big data resource management & networks: Taxonomy, survey, and future directions," *IEEE Communications Surveys & Tutorials*, 2021.
- [10] C. Stergiou, K. E. Psannis, B. B. Gupta, and Y. Ishibashi, "Security, privacy & efficiency of sustainable cloud computing for big data & IoT," *Sustainable Computing: Informatics and Systems*, vol. 19, pp. 174–184, 2018.
- [11] Q.-V. Pham, M. Zeng, R. Ruby, T. Huynh-The, and W.-J. Hwang, "UAV communications for sustainable federated learning," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 4, pp. 3944–3948, 2021.
- [12] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020.
- [13] J. Zhang, B. Chen, S. Yu, and H. Deng, "Peff: A privacy-enhanced federated learning scheme for big data analytics," in *2019 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2019, pp. 1–6.
- [14] A. Fu, X. Zhang, N. Xiong, Y. Gao, H. Wang, and J. Zhang, "VFL: A verifiable federated learning with privacy-preserving for big data in industrial IoT," *IEEE Transactions on Industrial Informatics*, 2020.
- [15] W. Y. B. Lim, N. C. Luong, D. T. Hoang, Y. Jiao, Y.-C. Liang, Q. Yang, D. Niyato, and C. Miao, "Federated learning in mobile edge networks: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 3, pp. 2031–2063, Third Quarter 2020.
- [16] D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li, and H. V. Poor, "Federated learning for internet of things: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, 2021.
- [17] M. Parimala, R. M. Swarna Priya, Q.-V. Pham, K. Dev, P. K. R. Maddikunta, T. R. Gadekallu, and T. Huynh-The, "Fusion of federated learning and industrial internet of things: A survey," *arXiv preprint arXiv:2101.00798*, 2021.
- [18] V. Mothukuri, R. M. Parizi, S. Pouriye, Y. Huang, A. Dehghantanha, and G. Srivastava, "A survey on security and privacy of federated learning," *Future Generation Computer Systems*, vol. 115, pp. 619–640, 2021.
- [19] L. Lyu, H. Yu, and Q. Yang, "Threats to federated learning: A survey," *arXiv preprint arXiv:2003.02133*, 2020.
- [20] J. Xu, B. S. Glicksberg, C. Su, P. Walker, J. Bian, and F. Wang, "Federated learning for healthcare informatics," *Journal of Healthcare Informatics Research*, vol. 5, no. 1, pp. 1–19, 2021.
- [21] J. Posner, L. Tseng, M. Aloqaily, and Y. Jararweh, "Federated learning in vehicular networks: opportunities and solutions," *IEEE Network*, vol. 35, no. 2, pp. 152–159, 2021.
- [22] G. Long, Y. Tan, J. Jiang, and C. Zhang, "Federated learning for open banking," in *Federated learning*. Springer, 2020, pp. 240–254.
- [23] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," *Mobile networks and applications*, vol. 19, no. 2, pp. 171–209, 2014.
- [24] B. Qolomany, A. Al-Fuqaha, A. Gupta, D. Benhaddou, S. Alwajidi, J. Qadir, and A. C. Fong, "Leveraging machine learning and big data for smart buildings: A comprehensive survey," *IEEE Access*, vol. 7, pp. 90 316–90 356, 2019.
- [25] L. D. Xu and L. Duan, "Big data for cyber physical systems in industry 4.0: a survey," *Enterprise Information Systems*, vol. 13, no. 2, pp. 148–169, 2019.
- [26] L. Zhu, F. R. Yu, Y. Wang, B. Ning, and T. Tang, "Big data analytics in intelligent transportation systems: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 1, pp. 383–398, 2018.
- [27] M. Ge, H. Bangui, and B. Buhnova, "Big data for internet of things: a survey," *Future generation computer systems*, vol. 87, pp. 601–614, 2018.
- [28] D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li, and H. V. Poor, "Federated learning for internet of things: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, pp. 1–1, 2021.
- [29] S. Abdulrahman, H. Tout, H. Ould-Slimane, A. Mourad, C. Talhi, and M. Guizani, "A survey on federated learning: The journey from centralized to distributed on-site learning and beyond," *IEEE Internet of Things Journal*, vol. 8, no. 7, pp. 5476–5497, 2021.
- [30] R. Patgiri, "Taxonomy of big data: A survey," *arXiv preprint arXiv:1808.08474*, 2018.
- [31] C.-H. Hua, T. Huynh-The, K. Kim, S.-Y. Yu, T. Le-Tien, G. H. Park, J. Bang, W. A. Khan, S.-H. Bae, and S. Lee, "Bimodal learning via trilogy of skip-connection deep networks for diabetic retinopathy risk progression identification," *International Journal of Medical Informatics*, vol. 132, p. 103926, 2019.
- [32] T. Huynh-The, C.-H. Hua, and D.-S. Kim, "Encoding pose features to images with data augmentation for 3-D action recognition," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 5, pp. 3100–3111, 2020.
- [33] N. A. Tu, T. Huynh-The, K. U. Khan, and Y.-K. Lee, "ML-HDP: A hierarchical bayesian nonparametric model for recognizing human actions in video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 3, pp. 800–814, 2019.
- [34] W. Yang, Y. Zhang, K. Ye, L. Li, and C.-Z. Xu, "Ffd: A federated learning based method for credit card fraud detection," in *International conference on big data*. Springer, 2019, pp. 18–32.
- [35] Y. Lu, X. Huang, Y. Dai, S. Maharjan, and Y. Zhang, "Blockchain and federated learning for privacy-preserved data sharing in industrial IoT," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 6, pp. 4177–4186, 2019.
- [36] L. Kong, X.-Y. Liu, H. Sheng, P. Zeng, and G. Chen, "Federated tensor mining for secure industrial internet of things," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 3, pp. 2144–2153, 2019.
- [37] Y. Lu, X. Huang, K. Zhang, S. Maharjan, and Y. Zhang, "Blockchain empowered asynchronous federated learning for secure data sharing in internet of vehicles," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 4, pp. 4298–4311, 2020.
- [38] C. Pappas, D. Chatzopoulos, S. Lalis, and M. Vavalis, "Ipls: A framework for decentralized federated learning," *arXiv preprint arXiv:2101.01901*, 2021.

- [39] M. Ghobakhloo, "Industry 4.0, digitization, and opportunities for sustainability," *Journal of cleaner production*, vol. 252, p. 119869, 2020.
- [40] J. Cheng, W. Chen, F. Tao, and C.-L. Lin, "Industrial IoT in 5G environment towards smart manufacturing," *Journal of Industrial Information Integration*, vol. 10, pp. 10–19, 2018.
- [41] M. H. ur Rehman, E. Ahmed, I. Yaqoob, I. A. T. Hashem, M. Imran, and S. Ahmad, "Big data analytics in industrial IoT using a concentric computing model," *IEEE Communications Magazine*, vol. 56, no. 2, pp. 37–43, 2018.
- [42] P. C. M. Arachchige, P. Bertok, I. Khalil, D. Liu, S. Camtepe, and M. Atiquzzaman, "A trustworthy privacy preserving framework for machine learning in industrial IoT systems," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 9, pp. 6092–6102, 2020.
- [43] S. Savazzi, M. Nicoli, and V. Rampa, "Federated learning with cooperating devices: A consensus approach for massive IoT networks," *IEEE Internet of Things Journal*, vol. 7, no. 5, pp. 4641–4654, 2020.
- [44] P. Zhang, C. Wang, C. Jiang, and Z. Han, "Deep reinforcement learning assisted federated learning algorithm for data management of IIoT," *IEEE Transactions on Industrial Informatics*, 2021.
- [45] S. Kim, "Incentive design and differential privacy based federated learning: A mechanism design perspective," *IEEE Access*, vol. 8, pp. 187317–187325, 2020.
- [46] L. C. FOURATI and A. Samiha, "Federated learning toward data pre-processing: Covid-19 context," in *2021 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 2021, pp. 1–6.
- [47] A. Imteaj and M. H. Amini, "Distributed sensing using smart end-user devices: pathway to federated learning for autonomous IoT," in *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*. IEEE, 2019, pp. 1156–1161.
- [48] K. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, C. Kiddon, J. Konečný, S. Mazzocchi, H. B. McMahan *et al.*, "Towards federated learning at scale: System design," *arXiv preprint arXiv:1902.01046*, 2019.
- [49] C. Zhang, Y. Xie, H. Bai, B. Yu, W. Li, and Y. Gao, "A survey on federated learning," *Knowledge-Based Systems*, vol. 216, p. 106775, 2021.
- [50] W. Xu, W. Fang, Y. Ding, M. Zou, and N. Xiong, "Accelerating federated learning for IoT in big data analytics with pruning, quantization and selective updating," *IEEE Access*, vol. 9, pp. 38457–38466, 2021.
- [51] R. Doku and D. B. Rawat, "Iflbc: On the edge intelligence using federated learning blockchain network," in *2020 IEEE 6th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*. IEEE, 2020, pp. 221–226.
- [52] Y. Wang, I. L. Bennani, X. Liu, M. Sun, and Y. Zhou, "Electricity consumer characteristics identification: A federated learning approach," *IEEE Transactions on Smart Grid*, 2021.
- [53] S. Zhai, X. Jin, L. Wei, H. Luo, and M. Cao, "Dynamic federated learning for gmec with time-varying wireless link," *IEEE Access*, vol. 9, pp. 10400–10412, 2021.
- [54] G. Xu, H. Li, S. Liu, K. Yang, and X. Lin, "Verifynet: Secure and verifiable federated learning," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 911–926, 2019.
- [55] J. So, B. Güler, and A. S. Avestimehr, "Byzantine-resilient secure federated learning," *IEEE Journal on Selected Areas in Communications*, 2020.
- [56] H. Li, D. Meng, H. Wang, and X. Li, "Knowledge federation: A unified and hierarchical privacy-preserving ai framework," in *2020 IEEE International Conference on Knowledge Graph (ICKG)*. IEEE, 2020, pp. 84–91.
- [57] Z. Ma, J. Ma, Y. Miao, X. Liu, K.-K. R. Choo, and R. Deng, "Pocket diagnosis: Secure federated learning against poisoning attack in the cloud," *IEEE Transactions on Services Computing*, 2021.
- [58] Y. Khazbak, T. Tan, and G. Cao, "Mlguard: Mitigating poisoning attacks in privacy preserving distributed collaborative learning," in *2020 29th International Conference on Computer Communications and Networks (ICCCN)*. IEEE, 2020, pp. 1–9.
- [59] Y. Shi, K. Davaslioglu, and Y. E. Sagduyu, "Over-the-air membership inference attacks as privacy threats for deep learning-based wireless signal classifiers," in *Proceedings of the 2nd ACM Workshop on Wireless Security and Machine Learning*, 2020, pp. 61–66.
- [60] S. Bhattacharya, S. R. K. Somayaji, T. R. Gadekallu, M. Alazab, and P. K. R. Maddikunta, "A review on deep learning for future smart cities," *Internet Technology Letters*, p. e187, 2020.
- [61] X. Xiao and C. Xie, "Rational planning and urban governance based on smart cities and big data," *Environmental Technology & Innovation*, vol. 21, p. 101381, 2021.
- [62] C. de Alwis, H. K. Arachchi, A. Fernando, and M. Pourazad, "Content and network-aware multicast over wireless networks," in *10th International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness*. IEEE, 2014, pp. 122–128.
- [63] E. Ismagilova, L. Hughes, N. P. Rana, and Y. K. Dwivedi, "Security, privacy and risks within smart cities: Literature review and development of a smart city interaction framework," *Information Systems Frontiers*, pp. 1–22, 2020.
- [64] Z. Zheng, Y. Zhou, Y. Sun, Z. Wang, B. Liu, and K. Li, "Applications of federated learning in smart cities: recent advances, taxonomy, and open challenges," *Connection Science*, pp. 1–28, 2021.
- [65] J. C. Jiang, B. Kantarci, S. Oktug, and T. Soyata, "Federated learning in smart city sensing: Challenges and opportunities," *Sensors*, vol. 20, no. 21, p. 6230, 2020.
- [66] B. Qolomany, K. Ahmad, A. Al-Fuqaha, and J. Qadir, "Particle swarm optimized federated learning for industrial IoT and smart city services," in *GLOBECOM 2020-2020 IEEE Global Communications Conference*. IEEE, 2020, pp. 1–6.
- [67] Y. Liu, A. Huang, Y. Luo, H. Huang, Y. Liu, Y. Chen, L. Feng, T. Chen, H. Yu, and Q. Yang, "FedVision: An online visual object detection platform powered by federated learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, New York, USA, 2020, pp. 13 172–13 179.
- [68] A. Albaseer, B. S. Ciftler, M. Abdallah, and A. Al-Fuqaha, "Exploiting unlabeled data in smart cities using federated edge learning," in *2020 International Wireless Communications and Mobile Computing (IWCMC)*. IEEE, 2020, pp. 1666–1671.
- [69] C. M. Thwal, K. Thar, Y. L. Tun, and C. S. Hong, "Attention on personalized clinical decision support system: Federated learning approach," in *2021 IEEE International Conference on Big Data and Smart Computing (BigComp)*. IEEE, 2021, pp. 141–147.
- [70] J. Pang, Y. Huang, Z. Xie, J. Li, and Z. Cai, "Collaborative city digital twin for the covid-19 pandemic: A federated learning solution," *Tsinghua Science and Technology*, vol. 26, no. 5, pp. 759–771, 2021.
- [71] P. Kumar, R. Kumar, G. Srivastava, G. P. Gupta, R. Tripathi, T. R. Gadekallu, and N. Xiong, "PPSF: A privacy-preserving and secure framework using blockchain-based machine-learning for IoT-driven smart cities," *IEEE Transactions on Network Science and Engineering*, 2021.
- [72] Y. Liu, X. Yuan, Z. Xiong, J. Kang, X. Wang, and D. Niyato, "Federated learning for 6g communications: Challenges, methods, and future directions," *China Communications*, vol. 17, no. 9, pp. 105–118, 2020.
- [73] W. Y. B. Lim, J. Huang, Z. Xiong, J. Kang, D. Niyato, X.-S. Hua, C. Leung, and C. Miao, "Towards federated learning in uav-enabled internet of vehicles: A multi-dimensional contract-matching approach," *IEEE Transactions on Intelligent Transportation Systems*, 2021.
- [74] N. Rieke, J. Hancox, W. Li, F. Milletari, H. R. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. A. Landman, K. Maier-Hein *et al.*, "The future of digital health with federated learning," *NPJ digital medicine*, vol. 3, no. 1, pp. 1–7, 2020.
- [75] S. Hakak, S. Ray, W. Z. Khan, and E. Scheme, "A framework for edge-assisted healthcare data analytics using federated learning," in *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 2020, pp. 3423–3427.
- [76] 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," *arXiv preprint arXiv:2101.07511*, 2021.
- [77] S. Silva, B. A. Gutman, E. Romero, P. M. Thompson, A. Altmann, and M. Lorenzi, "Federated learning in distributed medical databases: Meta-analysis of large-scale subcortical brain data," in *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*. IEEE, 2019, pp. 270–274.
- [78] Q. Wu, X. Chen, Z. Zhou, and J. Zhang, "Fedhome: Cloud-edge based personalized federated learning for in-home health monitoring," *IEEE Transactions on Mobile Computing*, 2020.
- [79] M. J. Sheller, B. Edwards, G. A. Reina, J. Martin, S. Pati, A. Kotrotsou, M. Milchenko, W. Xu, D. Marcus, R. R. Colen *et al.*, "Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data," *Scientific reports*, vol. 10, no. 1, pp. 1–12, 2020.
- [80] R. Wang, J. Xu, Y. Ma, M. Talha, M. S. Al-Rakhani, and A. Ghoneim, "Auxiliary diagnosis of covid-19 based on 5g-enabled federated learning," *IEEE Network*, vol. 35, no. 3, pp. 14–20, 2021.

- [81] I. Feki, S. Ammar, Y. Kessentini, and K. Muhammad, "Federated learning for covid-19 screening from chest x-ray images," *Applied Soft Computing*, vol. 106, p. 107330, 2021.
- [82] X. Li, Y. Gu, N. Dvornek, L. H. Staib, P. Ventola, and J. S. Duncan, "Multi-site fmri analysis using privacy-preserving federated learning and domain adaptation: Abide results," *Medical Image Analysis*, vol. 65, p. 101765, 2020.
- [83] O. El Rifai, M. Biotteau, X. de Boissezon, I. Megdiche, F. Ravat, and O. Teste, "Blockchain-based federated learning in medicine," in *International Conference on Artificial Intelligence in Medicine*. Springer, 2020, pp. 214–224.
- [84] R. Kumar, A. A. Khan, J. Kumar, A. Zakria, N. A. Golilarz, S. Zhang, Y. Ting, C. Zheng, and W. Wang, "Blockchain-federated-learning and deep learning models for covid-19 detection using ct imaging," *IEEE Sensors Journal*, 2021.
- [85] Z. Du, C. Wu, T. Yoshinaga, K.-L. A. Yau, Y. Ji, and J. Li, "Federated learning for vehicular internet of things: Recent advances and open issues," *IEEE Open Journal of the Computer Society*, vol. 1, pp. 45–61, 2020.
- [86] A. B. Sada, M. A. Bouras, J. Ma, H. Runhe, and H. Ning, "A distributed video analytics architecture based on edge-computing and federated learning," in *2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*. IEEE, 2019, pp. 215–220.
- [87] C. Xu and Y. Mao, "An improved traffic congestion monitoring system based on federated learning," *Information*, vol. 11, no. 7, p. 365, 2020.
- [88] Y. Qi, M. S. Hossain, J. Nie, and X. Li, "Privacy-preserving blockchain-based federated learning for traffic flow prediction," *Future Generation Computer Systems*, vol. 117, pp. 328–337, 2021.
- [89] Y. Liu, J. James, J. Kang, D. Niyato, and S. Zhang, "Privacy-preserving traffic flow prediction: A federated learning approach," *IEEE Internet of Things Journal*, vol. 7, no. 8, pp. 7751–7763, 2020.
- [90] N. I. Mowla, N. H. Tran, I. Doh, and K. Chae, "Federated learning-based cognitive detection of jamming attack in flying ad-hoc network," *IEEE Access*, vol. 8, pp. 4338–4350, 2019.
- [91] Y. M. Saputra, D. T. Hoang, D. N. Nguyen, E. Dutkiewicz, M. D. Mueck, and S. Srikanteswara, "Energy demand prediction with federated learning for electric vehicle networks," in *2019 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2019, pp. 1–6.
- [92] A. T. Thorgeirsson, S. Scheubner, S. Fünfgeld, and F. Gauterin, "Probabilistic prediction of energy demand and driving range for electric vehicles with federated learning," *IEEE Open Journal of Vehicular Technology*, vol. 2, pp. 151–161, 2021.
- [93] A. K. Bashir, S. Khan, B. Prabadevi, N. Deepa, W. S. Alnumay, T. R. Gadekallu, and P. K. R. Maddikunta, "Comparative analysis of machine learning algorithms for prediction of smart grid stability," *International Transactions on Electrical Energy Systems*, p. e12706, 2021.
- [94] A. Kumari, R. Gupta, S. Tanwar, S. Tyagi, and N. Kumar, "When blockchain meets smart grid: Secure energy trading in demand response management," *IEEE Network*, vol. 34, no. 5, pp. 299–305, 2020.
- [95] A. Taik and S. Cherkaoui, "Electrical load forecasting using edge computing and federated learning," in *ICC 2020-2020 IEEE International Conference on Communications (ICC)*. IEEE, 2020, pp. 1–6.
- [96] Y. L. Tun, K. Thar, C. M. Thwal, and C. S. Hong, "Federated learning based energy demand prediction with clustered aggregation," in *2021 IEEE International Conference on Big Data and Smart Computing (BigComp)*. IEEE, 2021, pp. 164–167.
- [97] J. Li, Y. Ren, S. Fang, K. Li, and M. Sun, "Federated learning-based ultra-short term load forecasting in power internet of things," in *2020 IEEE International Conference on Energy Internet (ICEI)*. IEEE, 2020, pp. 63–68.
- [98] M. Savi and F. Olivadese, "Short-term energy consumption forecasting at the edge: A federated learning approach," *IEEE Access*, 2021.
- [99] Q. Guo, F. Zhuang, C. Qin, H. Zhu, X. Xie, H. Xiong, and Q. He, "A survey on knowledge graph-based recommender systems," *IEEE Transactions on Knowledge and Data Engineering*, 2020.
- [100] A. Jalalirad, M. Scavuzzo, C. Capota, and M. Sprague, "A simple and efficient federated recommender system," in *Proceedings of the 6th IEEE/ACM International Conference on Big Data Computing, Applications and Technologies*, 2019, pp. 53–58.
- [101] B. Tan, B. Liu, V. Zheng, and Q. Yang, "A federated recommender system for online services," in *Fourteenth ACM Conference on Recommender Systems*, 2020, pp. 579–581.
- [102] C. Niu, F. Wu, S. Tang, L. Hua, R. Jia, C. Lv, Z. Wu, and G. Chen, "Billion-scale federated learning on mobile clients: a submodel design with tunable privacy," in *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, 2020, pp. 1–14.
- [103] K. Muhammad, Q. Wang, D. O'Reilly-Morgan, E. Tragos, B. Smyth, N. Hurley, J. Geraci, and A. Lawlor, "Fedfast: Going beyond average for faster training of federated recommender systems," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 1234–1242.
- [104] S. Niknam, H. S. Dhillon, and J. H. Reed, "Federated learning for wireless communications: Motivation, opportunities, and challenges," *IEEE Communications Magazine*, vol. 58, no. 6, pp. 46–51, 2020.
- [105] Y. Liu, J. Peng, J. Kang, A. M. Ilyasu, D. Niyato, and A. A. Abd El-Latif, "A secure federated learning framework for 5g networks," *IEEE Wireless Communications*, vol. 27, no. 4, pp. 24–31, 2020.
- [106] J. Zhou, S. Zhang, Q. Lu, W. Dai, M. Chen, X. Liu, S. Pirttikangas, Y. Shi, W. Zhang, and E. Herrera-Viedma, "A survey on federated learning and its applications for accelerating industrial internet of things," *arXiv preprint arXiv:2104.10501*, 2021.
- [107] A. K. Sangaiah, A. Goli, E. B. Tirkolaei, M. Ranjbar-Bourani, H. M. Pandey, and W. Zhang, "Big data-driven cognitive computing system for optimization of social media analytics," *IEEE Access*, vol. 8, pp. 82 215–82 226, 2020.
- [108] R. Doku, D. B. Rawat, and C. Liu, "Towards federated learning approach to determine data relevance in big data," in *2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI)*. IEEE, 2019, pp. 184–192.
- [109] 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. 53 841–53 849, 2020.
- [110] P. K. Sharma, J. H. Park, and K. Cho, "Blockchain and federated learning-based distributed computing defence framework for sustainable society," *Sustainable Cities and Society*, vol. 59, p. 102220, 2020.
- [111] D. J. Beutel, T. Topal, A. Mathur, X. Qiu, T. Parcollet, P. P. de Gusmão, and N. D. Lane, "Flower: A friendly federated learning research framework," *arXiv preprint arXiv:2007.14390*, 2020.
- [112] T. Ryffel, A. Trask, M. Dahl, B. Wagner, J. Mancuso, D. Rueckert, and J. Passerat-Palmbach, "A generic framework for privacy preserving deep learning," *arXiv preprint arXiv:1811.04017*, 2018.
- [113] C. He, S. Li, J. So, M. Zhang, H. Wang, X. Wang, P. Vepakomma, A. Singh, H. Qiu, L. Shen, P. Zhao, Y. Kang, Y. Liu, R. Raskar, Q. Yang, M. Annavaram, and S. Avestimehr, "Fedml: A research library and benchmark for federated machine learning," *arXiv preprint arXiv:2007.13518*, 2020.
- [114] S. Caldas, S. M. K. Duddu, P. Wu, T. Li, J. Konečný, H. B. McMahan, V. Smith, and A. Talwalkar, "Leaf: A benchmark for federated settings," *arXiv preprint arXiv:1812.01097*, 2018.
- [115] Y. LeCun, "The mnist database of handwritten digits," <http://yann.lecun.com/exdb/mnist/>, 1998.
- [116] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *CS224N project report, Stanford*, vol. 1, no. 12, p. 2009, 2009.
- [117] W. Shakespeare, *The complete works of William Shakespeare*. Wordsworth Editions, 2007.
- [118] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3730–3738.
- [119] T. Li, M. Sanjabi, A. Beirami, and V. Smith, "Fair resource allocation in federated learning," *arXiv preprint arXiv:1905.10497*, 2019.
- [120] V. Smith, C.-K. Chiang, M. Sanjabi, and A. Talwalkar, "Federated multi-task learning," *arXiv preprint arXiv:1705.10467*, 2017.
- [121] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.
- [122] P. M. Mammen, "Federated learning: Opportunities and challenges," *arXiv preprint arXiv:2101.05428*, 2021.
- [123] Q. Xia, W. Ye, Z. Tao, J. Wu, and Q. Li, "A survey of federated learning for edge computing: Research problems and solutions," *High-Confidence Computing*, p. 100008, 2021.
- [124] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings *et al.*, "Advances and open problems in federated learning," *arXiv preprint arXiv:1912.04977*, 2019.