Federated Learning and Blockchain

Subjects: Computer Science, Artificial Intelligence

Contributor: Muneerah Al Asqah , Tarek Moulahi

The Internet of Things (IoT) compromises multiple devices connected via a network to perform numerous activities. The large amounts of raw user data handled by IoT operations have driven researchers and developers to provide guards against any malicious threats. Blockchain is a technology that can give connected nodes means of security, transparency, and distribution. IoT devices could guarantee data centralization and availability with shared ledger technology. Federated learning (FL) is a new type of decentralized machine learning (DML) where clients collaborate to train a model and share it privately with an aggregator node.

Blockchain federated learning Internet of Things privacy

1. Federated Learning

Federated learning is a new technology that orchestrates connected clients to gain knowledge collaboratively. It was first introduced by Google in 2015 [1] to overcome three main issues: the huge amount of data gathered from many devices is unbalanced, non-independent and identically distributed (non-IID); the communication overhead of distant and massively distributed devices; and the insecure centralized data-storing mechanisms [2].

In federated learning, the learning burden is shared among connected nodes, usually referred to as clients, to train the ML model locally and upload the learning gradients to a central aggregator that levels all the learning gradients to a shared global model.

Figure 1 shows the basic topology of the federated learning procedure. It includes the list of clients selected to join the process by training the models locally. The aggregator is a central trusted server that could provide the aggregation results. Many communications can happen between different clients and the aggregator server. The process of FL is started by selecting clients, and next choosing the model. After performing the local training, the results will be sent to the aggregator server where they are aggregated. A global update can be performed next from the server and sent to the different clients. **Figure 1** can be described by these five essential steps:

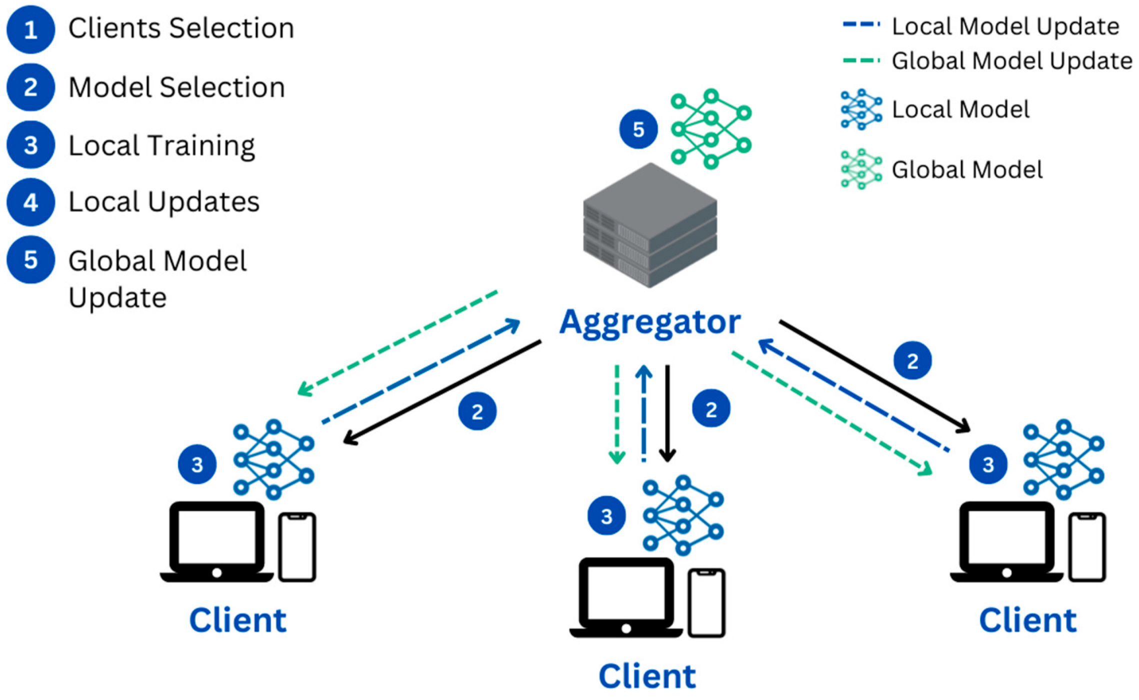

Figure 1. Basic topology of traditional federated learning.

- Clients selection: Participants' devices are selected to join the training iterations. This selection could depend on a number of factors, such as device processing capabilities and storage capacity, and is determined by definitive selection protocols [2].

- Model selection: The primary model is chosen, and its main parameters are determined and shared with clients to start the federated learning [3].

- Local model training: Clients independently train the model with the local device data storage [4].

- Local model gradients updates: After each iteration, clients push the training gradients to the aggregator device [2].

- Global model update: The aggregator applies an aggregation technique to level the trained model gradients and propagate the update to the clients to start the next round [4].

1.1. Categories of Federated Learning

In the literature, federated learning is categorized in two ways; one categorization is based on how the data are distributed [5], and the other is based on the network architecture [3]. The three data distribution categories are as follows:

- Horizontal federated learning: Where the datasets have the exact same features but varying samples.

- Vertical federated learning: Where the sample space is the same, but the features are different.

- Federated transfer learning: Starts from a pre-trained model where the overlap of the samples space and features space is less.

Based on how the devices are connected in the FL environment, it can be further classified as one of the following approaches:

- Centralized approach: Where a global central model is updated by aggregating the clients' training parameters. This approach applies protocols to avoid malicious client participation

- Decentralized approach: Where the clients' complete reliance on their neighbours to update the model removes the central authority. This approach requires absolute trust among clients.

1.2. Aggregation Techniques

Multiple algorithms are used to level the results from multiple participants' clients. **Table 1** summarizes three of the most used aggregation techniques. The researchers discuss the most relevant aggregation techniques, which are (1) FedAvg, which is based on calculating the parameters' average based on stochastic gradient descent (SGD); (2) SMC-Avg, which is characterized by its good performance even with 33% non-participated clients; and finally, (3) FedProx, which is derived from FedAvg, which can be applied in the case of heterogonous devices.

Table 1. Aggregation algorithms [6].

Algorithm	Based on	Centralized	Remarks
Federated Average (FedAvg)	Stochastic gradient descent (SGD)	√	-
Secure Multi-Party Computation (SMC-Avg)	-	√	Performs well even with 33% non-participating clients.
FedProx	FedAvg	√	Addresses device heterogeneity.

2. Blockchain Technology

Blockchain began after the publication of Nakamoto's white paper [7] on an electronic cash system. Although the term "Blockchain" was fairly new, the bundled technology consisted of cryptography and hashing mechanisms that were explored long before the Blockchain [8].

The definition of Blockchain is that it is a technology of peer-to-peer (P2P) networking that uses block-type data structures as storage, consensus mechanisms to manage a shared distributed ledger, and encryption to ensure security during data transmission [9].

Figure 2 illustrates an overview of Blockchain anatomy. Distributed ledger technology (DLT) includes a validated record of transactions in the form of blocks that contain a nonce value, transaction data, timestamp, and the previous block hash to form a chain, a Blockchain. This figure gives an overview of Blockchain architecture as a peer-to-peer (P2P) network. It is composed of a set of miners having each a copy of the Blockchain. A new block is added after the mining process, which is validated by at least by 51% of the miners. This makes Blockchain one of the most important systems in terms of protecting data integrity as well as transparency and availability.

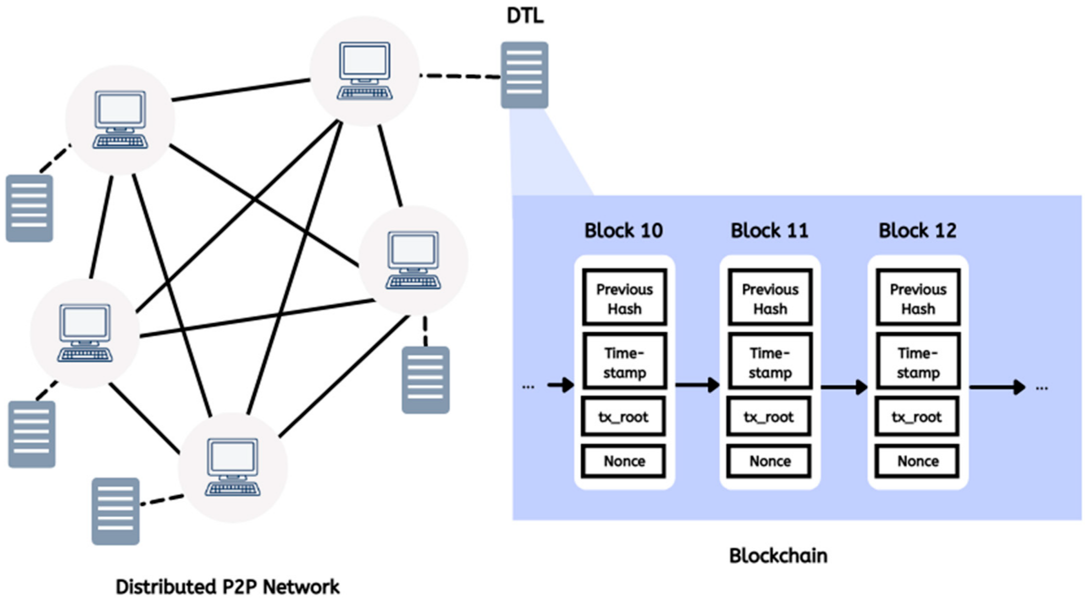

Figure 2. Blockchain as distributed DLT P2P technology.

2.1. Overview of Blockchain

Blockchain started with the first paper on Bitcoin in 2009 [7]. This era, the digital currency era, focused on developing decentralized-authority monetary transactional systems [10]. Next, research efforts was more focused on developing distributed applications (dApps) and the employment of smart contracts [9]. The use of artificial intelligence (AI) became integrated with Blockchain in order to be applied in industry 4.0 [11].

The type of Blockchain application is categorized based on its permissions as permissioned, permissionless, and federated Blockchain. Below, **Table 2** summarizes the differences between these three types [12]. In fact, there are three type of Blockchain. The type of the Blockchain can be defined based on four characteristics, which are whether the Blockchain is private or public and if it is controlled in a centralized or decentralized way. Two other characteristics that define the type of Blockchain are the level of security in addition to transaction speed and cost.

Table 2. Differences between permissionless, permissioned, and federated Blockchain [12].

Permissionless	Permissioned	Federated	
Publicity	Public	Private	Private
Authority	Decentralized	Centralized	Decentralized
Security	Less secure	Most secure	Secure
Transaction speed and cost	High	Less	Less

2.2. Components of Blockchain

Blockchain consists of multiple technology components which enable it to deliver the special characteristics of Blockchain, including security. There are six main Blockchain components which can be explained as follows:

- Cryptographic hash function: Blockchain employs hashing in two ways, in the cryptographic challenge and in the Merkle tree. The cryptographic challenge, the nonce, is the value that miner nodes compete to calculate. On the other hand, the Merkle tree is the representation of the transactions as hashed values [12].

- Asymmetric key encryption: Asymmetric encryption, or public-key encryption, is applied in addresses and digital signatures. The transactions are signed by the sender's private key, while the public key is used in the node's wallet address [12].

- Transactions: A transaction is the exchange of transmits, processes, and storages of digital assets to control the state among the Blockchain nodes. Several transactions will create a block.

- Consensus mechanisms: An agreement protocol to validate the new to-be-added block. Many consensus mechanisms exist. **Table 3** shows a brief review of the four most used and well-known consensus algorithms.

Table 3. Summary of consensus algorithms.

Consensus Algorithms	Steps	Blockchain	Remarks
Proof of Work [9]	• Transactions grouped into memory pool (mempool). • Miners try to solve the cryptographic challenge to validate. • The winner, the first to solve the challenge, is rewarded.	Public	First protocol in Blockchain [7]. High computational requirements. Less efficiency [11].

Consensus Algorithms	Steps	Blockchain	Remarks
	• Others verify the proof. A block (mempool) is attached.		
Proof of Stake [13]	• Nodes, validators invest an amount of stake (monetary value) to participate. • Random validator is selected. • Validator approves the block, gets rewarded. • If the block is malicious, validator is deprived of their stake.	Public	More resource efficient [9]. The selection is not that “random”. The higher a validator invests, the higher chance of being chosen [12].
Proof of Elapsed Time [9]	• Nodes wait for a random time. • After waiting, nodes become idle for a specific time. • The first to become active wins the block validation.	Private	System clock can be compromised [12].
Practical Byzantine Fault Tolerance	• A generator is chosen to collect and choose the block signors. • Signors use their digital signature to validate block integrity. • If the fault is *f*, 2*f* + 1 of 3*f* + 1 must reach a consensus.	Private	Addresses the scalability issues [9].

W), which was first used with Bitcoin. The Proof of Stake (PoS) was proposed to optimize the use of resources. The proof of elapsed time is a special type of consensus algorithm based on time. The fourth one is called practical Byzantine fault tolerance and addresses the scalability issues.

5. Smart contracts: It is a program that contains code and controls the state of the ledger through logic execution; if the conditions are met, the logic is invoked [14].

6. Ledger: The ledger contains the validated blocks and group of transactions. Others refer to it as the Blockchain memory [12].

2.3. Characteristics of Blockchain

The properties are the decentralization behavior, transparency, immutability and traceability, trusting, and anonymity. The numerous Blockchain components enable it to possess the following features [9][14]:

- Decentralization: where the ledger is shared among all the P2P network nodes.
- Transparency: where the ledger records are retrievable by any Blockchain node.
- Immutability and traceability: Where each block points to its predecessor, meaning a change to one block's content will not go unnoticed. Furthermore, where each block is timestamped to enhance the data traceability.
- De-Trusting: where no central authority or a third party is required to review the operations.
- Anonymity: where nodes are identified by their digital signature.
- Credibility: where internal calculations are automatically performed without human intervention, making Blockchain credible to perform secure operations.

3. Taxonomy of Federated Learning, Blockchain, and IoT

The integration of these two powerful technologies has many applications. Deploying Blockchain and federated learning integration in the IoT with all its different environments has become the main direction of authors and researchers [3]. **Figure 3** illustrates the basic taxonomy of all layers of these technologies [2]. The Blockchain technology is foreseen to create a revolution in both industry and commerce, making great global economic changes, as it is immutable, transparent, and redefines trust, offering secure, fast, reliable, and transparent solutions. The IoT can leverage the absence of intermediates in the Blockchain, enabling users to communicate directly with IoT devices with no one intercepting them, which could offer a huge application area [15]. The following four layers of architecture can describe the general framework to develop application- and solution-merging between IoT, FL, and Blockchain. The IoT is responsible for data collection. The AI process is applied through FL for privacy-preservation issues in case of sensitive data such as patient data. The aggregation, finally, is performed in the Blockchain as a trusted and confident layer.

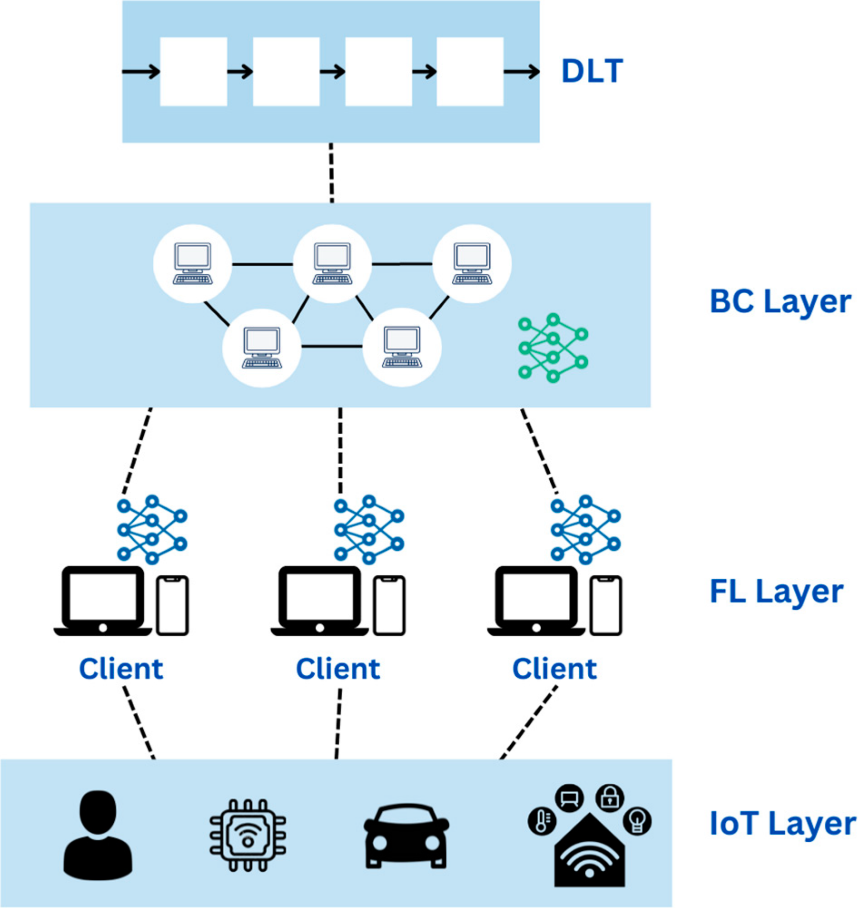

Figure 3. Taxonomy of federated learning, Blockchain, and IoT.

This integration between federated learning and Blockchain in the IoT could be shaped differently according to the type of application. Researchers of [5] provided a clear explanation of this integration's different architectures. However, it is out of this paper's scope.

References

1. McMahan, H.B.; Moore, E.; Ramage, D.; Hampson, S.; Arcas, B.A.Y. Communication-Efficient Learning of Deep Networks from Decentralized Data. arXiv 2017.
2. Wang, Z.; Hu, Q. Blockchain-based Federated Learning: A Comprehensive Survey. arXiv 2021, arXiv:2110.02182.
3. Ali, M.; Karimipour, H.; Tariq, M. Integration of blockchain and federated learning for Internet of Things: Recent advances and future challenges. Comput. Secur. 2021, 108, 102355.
4. Konečný, J.; McMahan, H.B.; Yu, F.X.; Richtárik, P.; Suresh, A.T.; Bacon, D. Federated Learning: Strategies for Improving Communication Efficiency. arXiv 2017, arXiv:1610.05492.

5. Yang, Q.; Liu, Y.; Chen, T.; Tong, Y. Federated Machine Learning: Concept and Applications. arXiv 2019.
6. Issa, W.; Moustafa, N.; Turnbull, B.; Sohrabi, N.; Tari, Z. Blockchain-based federated learning for securing internet of things: A comprehensive survey. ACM Comput. Surv. 2023, 55, 1–43.
7. Nakamoto, S. Bitcoin: A Peer-to-Peer Electronic Cash System. p. 9. Available online: https://www.ussc.gov/sites/default/files/pdf/training/annual-national-training-seminar/2018/Emerging_Tech_Bitcoin_Crypto.pdf (accessed on 28 October 2022).
8. Haber, S.; Stornetta, W.S. How to time-stamp a digital document. J. Cryptol. 1991, 3, 99–111.
9. Namasudra, S.; Deka, G.C.; Johri, P.; Hosseinpour, M.; Gandomi, A.H. The Revolution of Blockchain: State-of-the-Art and Research Challenges. Arch. Comput. Methods Eng. 2021, 28, 1497–1515.
10. Efanov, D.; Roschin, P. The All-Pervasiveness of the Blockchain Technology. Procedia Comput. Sci. 2018, 123, 116–121.
11. Cummings, S. The Four Blockchain Generations. The Capital. 2 February 2019. Available online: https://medium.com/the-capital/the-four-blockchain-generations-5627ef666f3b (accessed on 22 November 2022).
12. Yaga, D.; Mell, P.; Roby, N.; Scarfone, K. Blockchain Technology Overview; NIST Internal or Interagency Report (NISTIR) 8202; National Institute of Standards and Technology: Gaithersburg, MA, USA, 2018.
13. Proof-of-Stake (PoS). Ethereum.Org. Available online: https://ethereum.org (accessed on 23 November 2022).
14. Lu, Y. The blockchain: State-of-the-art and research challenges. J. Ind. Inf. Integr. 2019, 15, 80–90.
15. Alfrhan, A.; Moulahi, T.; Alabdulatif, A. Comparative study on hash functions for lightweight blockchain in Internet of Things (IoT). Blockchain Res. Appl. 2021, 2, 100036.

Retrieved from https://encyclopedia.pub/entry/history/show/109937