Self-Sovereign Identity Management for Hierarchical Federated Learning in Vehicular Networks

Engin Zeydan*, Josep Mangues*, Suayb S. Arslan†, Yekta Turk◇
*Centre Tecnològic de Telecomunicacions de Catalunya (CTTC), Barcelona, Spain, 08860.
† Massachusetts Institute of Technology, MA, USA, 02139.
◇Mobile Network Architect, Istanbul, Turkey, 34396.
Email: engin.zeydan@cttc.cat, sarslan@mit.edu, yektaturk@gmail.com

Abstract—**There has been a rapid increase in the number of connected vehicles with a huge amount of data exchange between these vehicles that needs to be communicated, processed and analyzed reliably and efficiently. For secure and decentralized authentication, self-sovereign identity (SSI) management in vehicular networks have attracted attention in recent years. Hierarchical deployment frameworks, on the other hand, can provide secure and efficient knowledge sharing for vehicular networks with heterogeneous and geographically distributed vehicles and infrastructure in 6G networks. In this paper, we explore the joint use of hierarchical federated learning, as a collaborative machine learning framework, and hierarchical SSI management in vehicular networks, highlighting its advantages, limitations. At the end of the paper, we also provide two illustrative use cases.**

Keywords*—*self-sovereign, digital identity, blockchain, hierarchical federated learning, vehicular networks.

I. Introduction

6G is expected to provide low-delay communication, dependable connections, and advanced sensory abilities to vehicular networks. This helps increase connectivity and provide novel Artificial Intelligence (AI) and Machine Learning (ML)-empowered vehicular applications for their users at the edge. Future vehicle communication networks are seeking cheap and efficient vehicle to vehicle (V2V), vehicle to infrastructure (V2I), and vehicle to everything (V2X) solutions mostly relying on Intelligent Transportation System (ITS) that utilize the power of AI/ML-based approaches for road-safety, emergency responses, traffic optimization, road optimization, or vehicle maintenance.

Learning in vehicular networks can manage the available resources smartly and efficiently. However, it should cover large geographical regions with different road traffic dynamics, and network characteristics. Combining such diverse data and pooling the model parameters for secure and distributed learning becomes one of the most challenging tasks. Therefore, conventional Federated Learning (FL) techniques fail to address this issue as they require a fairly large number of resources for data transmission and storage. Hierarchical Federated Learning (HFL) has emerged as a promising approach for addressing this challenge, as it allows for decentralized learning and decision-making while still preserving privacy and data security. HFL is a type of FL where the data is organized into a hierarchy of groups or clusters. In HFL, each group has its own local model that is trained on the data within its cluster. Later at the top level of the hierarchy, the models are combined to form a global model [1]. These groups can be vehicles as well as infrastructures (Roadside Units (RSUs) and Base Stations (BSs)) that are distinguished according to their regional and functional features.

Together with the increasing connectivity of vehicles and the growing amount of sensitive information being transmitted and stored in these networks, it is crucial to have a robust and scalable solution for identity management. Blockchain-based Self-Sovereign Identity (SSI) provides decentralized identity management and allows vehicle users to have full control and ownership over their own personal data (e.g., selectively sharing information with privacy control) [2] while also providing a secure authentication and authorization process. This will enable the development of secure, privacy-preserving, and highly automated road transportation networks. In SSI systems, Blockchain Networks (BCNs) help identities to be created, stored, and managed, so that the data is secured and cannot be tampered with or

This work was partially funded by "ERDF A way of making Europe" project PID2021-126431OB-I00 and Spanish MINECO - Program UNICO I+D (grants TSI-063000-2021-54 and -55) Grant PID2021-126431OB-I00 funded by MCIN/AEI/ 10.13039/501100011033.

misused. It also allows for the verification of identities, secure storage and transmit of data, and provide audit trails, allowing for easier tracking of any changes that are made to the ledger.

II. Related Work and Our Contributions

The 5G NR standard has already new use cases for Cellular Vehicle-to-Everything (CV2X) communications such as remote driving, vehicle platooning, extended sensors, and advanced driver assistance functionalities [3]. 6G use cases are expected to go beyond in V2X communication with technologies such as brain–vehicle interfacing, tactile communication, and terahertz (THz) communications [4] and use cases such as Advanced remote driving, intelligent traffic scheduling, holographic driving, personalized transport vehicles with holographic infotainment and tactile/haptic interactions, flying taxis and cargo drones [5]. FL-based frameworks for decentralized training of ML models in vehicular networks already exist in the literature [6]. HFL is also used for vehicles to learn environmental data and share the learning knowledge with each others while relying on BCN ability to deal with certain malicious attacks effectively by authors in [7].

To provide more secure and dependable digital identification solutions, SSI-based smart contract implementations such as Jolocom[1], OpenSSI, and Sovrin are being developed. To ensure the privacy and security of users' personal data, these solutions employ distributed ledger technology and smart contracts. They also give a safe and dependable method for users to authenticate their identification to third parties without disclosing any sensitive information. This gives users control over who has access to their data and how it is utilized. Organizations may utilize these solutions to build trustless and secure digital identification systems that can be used for a number of purposes. A comprehensive review and mapping of theoretical and practical advances in SSI is provided in [8]. The authors in [9] review use cases, technologies and challenges of SSI within Industrial Internet of Things (IIoT) applications.

Considering the above related works, to the best of authors' knowledge, no work has been done exploiting the benefits of SSI during HFL process, which forms the basis of our motivation in our work. Our main contributions are as follows: (i) Combining the benefits of hierarchical SSI in a HFL system for vehicular networks, (ii) Providing a novel mathematical formulation for the SSI embedded HFL, (iii) Describing the benefits of the proposed approach in two novel use cases. The rest of the paper is organized as follows: Section III discusses about the essentials of HFL, its potential in vehicular networks and the role of BCN. Section IV discusses about providing security in identity management and how HFL can be combined with SSI. Section V gives the phases of the proposed HFL with SSI solution. Section VI discusses about the considered use case and the potential future enhancements to the secure architecture. Finally, Section VII gives the conclusions of the paper.

III. HFL in vehicular networks

A. FL and HFL Essentials

We aim at learning a unique global statistical model via supervision at a central server, whereby each of the N workers possess a private dataset and helps with the local model training. Let us denote the private dataset owned by i-th worker (client) by $\mathcal{D}^i = \{\mathbf{x}_j^{(i)}, y_j^{(i)}\}$, where $i \in \{1, 2, \ldots, N\}$, $\mathbf{x}_j^{(i)}$ is the j-th input sample from $\mathcal{D}^i$ and $y_j^{(i)}$ is the label information. Let us also denote $\boldsymbol{\ell}(\mathbf{x}_j^{(i)}, y_j|\mathbf{w})$ as the loss function that is used to minimize the empirical loss $\mathcal{L}(\mathbf{w})$, where $\mathbf{w}$ is the hyperparameter set of the global model. The main objective of FL is to solve the following optimization problem using distributed training,

$$\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}) = \frac{\sum_{i=1}^{n} \sum_{j \in \mathcal{D}^i} \boldsymbol{\ell}(\mathbf{x}_j, y_j|\mathbf{w})}{|\mathcal{D}_\cup|} \tag{1}$$

where $\mathcal{D}_\cup = \cup_{i=1}^{n} \mathcal{D}^i$. For faster convergence and efficient computation, it is typical to use mini-batch sthocastic gradient descent (SGD) (with batch B of size m) to find the minimum cost where, in each step, model parameters are updated as,

$$\mathbf{w} \leftarrow \mathbf{w} - \frac{\alpha}{m} \sum_{j \in B} \frac{\partial \boldsymbol{\ell}(\mathbf{x}_j, y_j|\mathbf{w})}{\partial \mathbf{w}} \tag{2}$$

where α is the learning rate scaled by the average of the mini-batch gradients through back-propagation. Each device i performs its mini-batch SGD on their corresponding dataset $\mathcal{D}^i$ to find the local model parameters $\mathbf{w}_i$. Finally, the central server aggregates all local parameters to update the global ones which is shared in the next step with workers for convergence.

HFL allows multiple levels of servers and devices to be used to aggregate and distribute the machine learning models, typically in a star topology. This approach allows

[1] Online: https://jolocom.io/, Available: January 2023.

for more efficient communication, better scalability, and more robustness to device failures or disconnections. It also enables different levels of privacy and security to be applied to the data and models at different levels of the hierarchy. The central server typically runs a clustering of workers and assigns an aggregator (an edge server) from each cluster that establishes up-and-down links between the workers and the central server. Since the aggregators serve as the intermediary (a relay in a downstream transmission) and lump together all the collected model weights (also hyperparameters), it helps with the overall network traffic and bandwidth. It reduces the amount of data that needs to be transmitted between clients and the server since all the model weights are collected in one spot. This helps to reduce latency and improve the speed of training.

B. Benefits of HFL for Vehicular Scenarios

In vehicular networks, HFL can be very useful in various applications to manage vehicle flow or traffic over a group or clusters of users. For example, 3D object detection by utilizing image datasets, feature learning in different geographical locations, collaborative driving with lane changing, semantic segmentation for self-steering and route optimization for fuel efficiency are only few to name. In each of these objectives pertaining to vehicular network optimizations, the data is collected from the vehicles and then shared among the edge servers or other vehicles that serve as cluster parameter servers in the network. Each dataset is then used to train a HFL model, which is then used to control the behavior of the vehicles in the network through dissemination of the learned model to each user of the system. This can help improve the efficiency of the running traffic, reduce congestion, and improve fuel efficiency.

Current static topologies considered for HFL focus on latency and bandwidth resorces and may not be directly applicable to the requirements of vehicular networks. As the users of the network are mobile, user clusters might be forming continuously to accommodate the ever-changing cluster size, node locations, and user dataset distributions. The convergence of the overall model training is therefore a function of the mobility and the iteration efficiency of the global and local model parameter exchanges.

C. Storage of Model Updates: Blockchain Network

In HFL, due to local/global updates, there appear various model states that follow a specific trajectory. Tracing the training procedure (model updates) can reveal many properties of the overall learning process, potential biases in local datasets housed by the system workers. For temper-proof tracing, Blockchain may be used within the context of HFL to organize and store decentralized model updates that are produced by participants at various levels of the hierarchy on a secure/open platform. The model updates are recorded irreversibly due to the secure decentralized nature of blockchain, which also protects the system from harmful assaults and unintended data manipulations. Additionally, the HFL system's rules and procedures, such as rewarding members who contribute to the model update the most useful way or settling any conflicts that may develop during the training process, may be handled by *smart contracts*. Overall, implementing blockchain in HFL enhances the system's security, dependability, and scalability.

IV. Fusion of SSI within HFL Framework

A. Security in Identity Management

Self-sovereign identity management (SSI) is a digital identification strategy that prioritizes user control, privacy, and security. It enables individuals to own and control their personal identifying data, as well as share it selectively with others, without the need for intermediaries or centralized authority. SSI systems use decentralized, distributed ledger technologies (such as blockchain) and provide safe, private, and portable identity credentials that may be utilized across multiple applications and services. SSI can be structured in a hierarchical manner in which each level of the hierarchy can issue and verify credentials for the entities below it, while also relying on the credentials issued by the levels above it. This creates a chain of trust that extends from the root of the hierarchy to the individual entities at the bottom.

As the number of connected devices and the transmission of sensitive information grows, security will be a major concern for cellular technology. Effective security measures such as secure device identification, safe data storage, and encrypted data transmission will be crucial for protecting the privacy and security of users and their data. As a solution to device authentication, hierarchical SSI can allow each individual device, user, etc. to have a digital identity within a cluster and control over their personal data while relying on BCN. They can also selectively share it with different organizations or persons at different levels of trust and verifiability. As a result, users can safely access and keep their own data while, at their choice, granting access to other people or organizations. Users can also establish digital identities that are connected to their physical identities. This enables them to safely access the services and resources that are made available by other users or companies. Users may now access, store, and share their data in a safe manner while maintaining their privacy and security.

B. A HFL Framework Leveraging SSI

Unlike in the conventional HFL where workers are the sole owners of their dataset, in a combined HFL and hierarchical SSI scenario, multiple entities can be authorized to access datasets, leading to a hybrid sharing of both model parameters as well as datasets or their compressed representations. Hierarchical SSI ensures the data (which can be organized into groups based on trust level) to be shared only with enterprises or persons that are trusted while ensuring them control over their datasets. This can be performed using digital signatures and encryption algorithms that can be tied to the SSI of each entity. HFL, on the other hand, ensures the model training done on the private data within each secure group. By combining HFL with hierarchical SSI, model training can be done on high-quality and personal data within each group while preserving the individual privacy and control over such private data across HFL secure groups. The strategies might differ depending on how the hierarchical SSI and HFL groups are defined. In fact in the proposed scheme each worker is associated with two groups, one virtually defines the channel over which parameter sharing is done using HFL, whereas the other defines how the additional data sharing is enabled within the same network of workers using SSI.

C. Objective Subject to SSI

Let us assume $\mathcal{N}$ HFL clusters and $\mathcal{M}$ HSSI groups which share the same pool of workers. For a given cluster $i \in \{1, \ldots, \mathcal{N}\}$, we have i_j number of nodes sharing the same privacy hierarchy $j \in \{1, \ldots, \mathcal{M}\}$ within the SSI context. Assuming the set $\mathcal{R}^{ij}$ to be a representation of all the private datasets of i_j nodes/workers in the cluster i and privacy hierarchy j. Since the data (or its representation) is shareable among these i_j nodes subject to *SSI creation* (provided in the next section), learning can be totally asynchronous, decentralized and offline. Additionally, to mitigate the deviations of the local models on an arbitrary SSI hierarchy j from that of the central server, a regularized loss function is minimized [10] instead, namely,

$$\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}) = \frac{1}{A}\sum_{i=1}^{|\mathcal{N}|}\sum_{j=1}^{|\mathcal{M}|}\sum_{k\in\mathcal{R}^{ij}} \ell(\mathbf{x}_k, y_k|\mathbf{w}_k) + \frac{\lambda}{2}||\mathbf{w}-\mathbf{w}_k||^2$$

where $A = 1/\sum_i \sum_j i_j$, each sum in the numerator

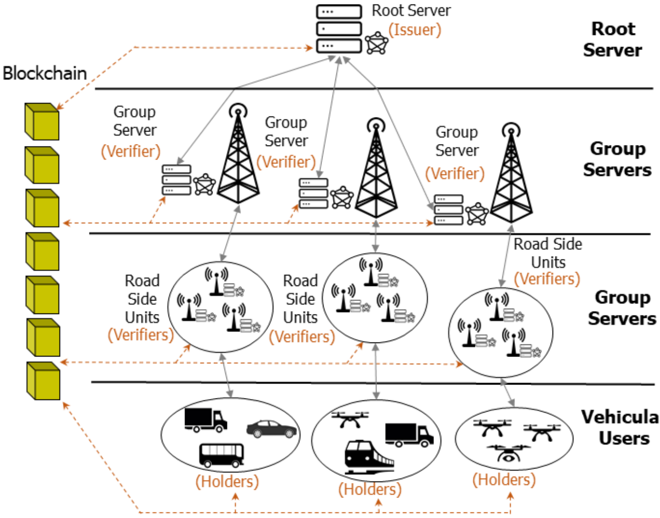


Fig. 1: BCN based SSI framework for hierarchical federated learning in vehicular networks.

demonstrates a hierarchy in the FL and $\frac{\lambda}{2}||\mathbf{w}-\mathbf{w}_k||^2$ is the regularization term to control the deviation of local and central model parameters. As can be seen, iterations are now conducted between hierarchies of the FL (BSs, RSUs and workers) as well as across different hierarchical SSI groups.

V. Phases of the Proposed Solution

The process of combining HFL and hierarchical SSI may involve participation of several entities and may require several iterations of implementation phases. As shown in Fig. 1, an initial local FL training is done among vehicles and RSUs in a small region. Then, a higher level FL process is performed among RSUs and BSs in a larger geographical area. By employing HFL, different features belonging to various traffic regions can be accommodated. HFL using hierarchical SSI solutions can be described as follows:

Step-0 *(Identification of Hierarchy):* Both HFL and SSI need a clear definition of the hierarchy among the participating nodes/workers. In our analysis, we have identified three main roles and responsibilities: vehicles (as data sources), RSUs (as both data provider and aggregator), and BSs (as top-level (central) data aggregators).

Step-1 *(SSI Creation):* In this step, each entity involved in the learning process needs to establish its SSI first, and each entity's owned SSIs is created and linked with their personal data using a decentralized management solution based on BCNs.

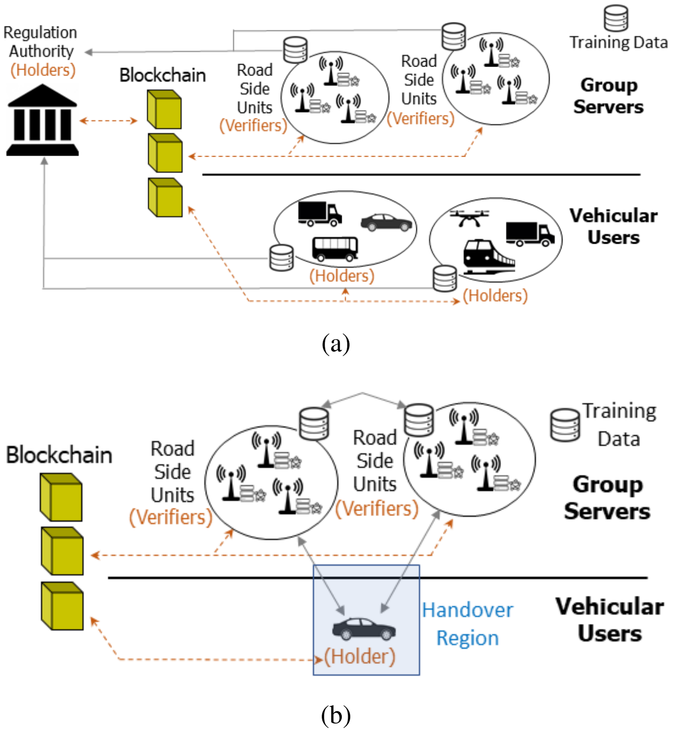
Step-2 *(Local Model Training):* Each vehicle user in the bottom layer trains their local model to refine hyperparameters based on their own private data.

STEP-3 *(Sharing Parameter Learning):* In this step, organizations or vehicle user requests access to the parameters of the personal vehicle model learning and the owner of the personal data decides whether or not to share it with them. This decision depends on the trust level established between the vehicle user and the organization requesting the data.

STEP-4 *(Hierarchical Model Organization):* The shared learning parameters by the vehicle users are organized into groups depending on the trust level. These shared learning results are recorded in BCN. An exclusive distributed ledger is used to enhance security during the sharing process. The interaction between vehicles during HFL and hierarchical SSI process is encapsulated as a form of transaction. This transaction can be audited and recorded by all peers in the network via a consensus protocol for immutable storage. The learning results are sent to nearby servers (e.g., RSUs) which collect transactions within their reach and build blocks in BCN (through consensus within the same secure group) which typically contain trained parameters of the local FL.

STEP-5 *(Model Aggregation):* The trained model parameters within each group are aggregated at the top-level hierarchy to obtain a global model. In the middle layer, RSUs work as clients in the global FL and integrate results from the workers layer (which are obtained from vehicle clients) and its own learning parameters built based on the sensing environment to train middle layer learning parameters. Similar to the previous step, in this layer, BSs collect the transactions sent by RSUs and assembles them into blocks in BCN. In this step, the ledger contains the shared knowledge from both RSUs and vehicle users that participated in FL. Note that during this process, RSUs have two different digital identities, i.e., worked both as publishers of blocks in lower layer BCN and the producers of transactions in top layer BCN.

STEP-6 *(Model Distribution):* Global model parameters are distributed back (relayed) to all vehicle users in the bottom layer. This model can be used for various purposes such as image detection, traffic analysis, etc. from a global perspective while keeping each vehicle user preserve and control their personal data. We finally remark that during the above 6-step process, all vehicles, RSUs and BSs work collaboratively to obtain a global shared model while ensuring the privacy as well as immutable persistence of the generated data among constituent secure groups of the system.



(a)

(b)

Fig. 2: SSI based identities can be configured for various use cases in vehicular networks with HFL, (a) Road safety, (b) Data sharing for handover.

VI. USE CASES AND SECURITY ANALYSIS

Two use cases as described in Fig. 2 are given in this section. **(i) Road Safety:** Combining HFL with SSI can improve the road safety of vehicular networks through providing a secure, decentralized and privacy-preserving solution as given in Fig. 2(a). In a HFL setting, vehicles can collect data from various sensors including Global Positioning System (GPS), sensors, accelerometers or cameras, whereas RSUs can collect their own data (e.g., long historical data, critical data). The collected data from RSUs and vehicles based on the hierarchy can be used to train ML models using HFL to detect and predict potential hazards in the road real-time. This enables training process to be concluded without having to share any sensitive data with a central server. Hierarchical SSI, on the other hand, can be used to prevent unauthorized access to sensitive information and ensure secure data exchange with BCN while registration phase of HFL process. In case of erroneous outcomes of the model prediction, regulation authority can retain the control over the top level data and can delegate access to lower level data. **(ii) Data sharing for handover:** In this use case, the data can be partitioned into different levels, with the most sensitive data at the top level (such as vehicle user information) and less sensitive data (such as vehicle location, speed or network conditions) at bottom levels as given in Fig. 2(b). Together with SSI-enabled HFL, multiple vehicles can be allowed to share information about the network conditions and optimize the handover process in the cellular network. Finally, in both use cases, HFL can be implemented using existing FL frameworks such as TensorFlow Federated[2] or PySyft[3].

The SSI-based solution stores the verification and encryption keys on the blockchain while the signing and decryption keys remain on the HFL nodes. This offers defense against serious cyberattacks including phishing, man-in-the-middle attacks and replay attacks for the locally trained ML models. A public, unchangeable blockchain ledger ensures the system's security and the veracity of the constituted model. Secure access management, a security element of the blockchain authentication protocol, will help reduce the need for time-consuming and expensive enterprise-wide password reset procedures. An identity management system's complexity increases the number of potential attack surfaces that are used to compromise system security and steal data. By using blockchain helps to keep this attack surface small. Identity management records (including creation, deletion, and update) are included in the ledger, and the integrity of the data is clear. For identity management and forensic reasons, secure blockchain authentication ledgers offer immutable data.

In addition to the temper-proof model update storage, smart contract-based incentives can be used to help participants to contribute to the overall learning process. For example, rewards can be automatically distributed to the participants who contribute to the model updates. Participants in the HFL system can communicate their data with one another without disclosing their raw data by using privacy-preserving technologies like *homomorphic encryption*. This aids in preserving the participants' and their data's privacy. Connected to privacy, with SSI implemented on blockchain, participants can have full control over their personal data, thereby reducing the risk of data breaches and misuses. BCNs may also be applied to provide a decentralized forum for settling potential participant conflicts. This can support maintaining the system's fairness and openness. Plus, the resistance to quantum computers can be obtained by using post-quantum security techniques [11].

VII. Conclusions

Using SSI during HFL is a promising approach and can provide a secure and decentralized system for training machine learning models hierarchically in vehicular networks. It can also help maintain the privacy and autonomy of each participating entity in the vehicular network (i.e., vehicles, RSUs or BSs). In this paper, we investigated how blockchain-based SSI integration can provide a robust and scalable solution to identity management in vehicular networks during HFL. Incorporating SSI in the optimization formulations of HFL phases is briefly described. With potential 6G applications, two novel use cases are also proposed that can show the potential of the proposed approach to improve the security and efficiency of vehicular networks. A few limitations and challenges such as ensuring fairness during HFL process or scaling in complex and large vehicular network scenarios for future 6G use cases still need to be addressed.

References

[1] L. Liu *et al.*, "Client-edge-cloud hierarchical federated learning," in *IEEE Int. Conf. on Communications (ICC)*, pp. 1–6, IEEE, 2020.

[2] M. R. Ahmed *et al.*, "Blockchain-based identity management system and self-sovereign identity ecosystem: A comprehensive survey," *IEEE Access*, vol. 10, pp. 113436–113481, 2022.

[3] S. A. Abdel Hakeem *et al.*, "5G-V2X: Standardization, architecture, use cases, network-slicing, and edge-computing," *Wireless Networks*, vol. 26, pp. 6015–6041, 2020.

[4] M. Noor-A-Rahim *et al.*, "6G for vehicle-to-everything (V2X) communications: Enabling technologies, challenges, and opportunities," *Proceedings of the IEEE*, vol. 110, no. 6, pp. 712–734, 2022.

[5] V.-L. Nguyen *et al.*, "Towards the age of intelligent vehicular networks for connected and autonomous vehicles in 6G," *IEEE Network*, 2022.

[6] A. M. Elbir *et al.*, "Federated learning in vehicular networks," in *2022 IEEE Int. Mediterranean Conf. on Communications and Networking (MeditCom)*, pp. 72–77, IEEE, 2022.

[7] H. Chai *et al.*, "A hierarchical blockchain-enabled federated learning algorithm for knowledge sharing in internet of vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 7, pp. 3975–3986, 2020.

[8] F. Schardong and R. Custódio, "Self-sovereign identity: A systematic review, mapping and taxonomy," *Sensors*, vol. 22, no. 15, p. 5641, 2022.

[9] P. C. Bartolomeu *et al.*, "Self-sovereign identity: Use-cases, technologies, and challenges for industrial IoT," in *2019 24th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, pp. 1173–1180, IEEE, 2019.

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

[11] E. Zeydan, Y. Turk, B. Aksoy, and Y. Y. Tasbag, "Post-quantum era in V2X security: Convergence of orchestration and parallel computation," *IEEE Communications Standards Magazine*, vol. 6, no. 1, pp. 76–82, 2022.

[2] Online: https://www.tensorflow.org/federated, Available: February-2023.

[3] Online: https://github.com/OpenMined/PySyft, Available: February-2023.