Blockchain-Based Self-Sovereign Identity for Federated Learning in Vehicular Networks

Engin Zeydan*, Luis Blanco*, Josep Mangues*, Suayb 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, luis.blanco, josep.mangues}@cttc.cat, sarslan@mit.edu, yektaturk@gmail.com

***Abstract*—Self-Sovereign Identity (SSI) has emerged lately as an identity and access management framework that is based on Distributed Ledger Technology (DLT) and allows users to control their own data. Federate Learning (FL), on the other hand, provides a framework to update Machine Learning (ML) models without relying on explicit data exchange between the users. This paper investigates identity management and authentication for vehicle users, which are participating into FL. We propose a new approach to SSI, that is alternative to the conventional blockchain-based SSI, specifically for use in vehicular networks, which focuses on maintaining confidentiality, authenticity, and integrity of vehicle users' identities and data exchanged between the users and the aggregation server during the execution of the FL process. We also provide experimental results for distributed identity management (DIM) operations, which show that the performance of credential operations in the implemented system is generally efficient and the average times are within reasonable limits. However, there is a slight increase in presentation time, offer time, connection establishment time, and credential revocation time as the number of requests increases, indicating a slight degradation in performance for these operations.**

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

I. Introduction

Web 3.0 has received increasing attention which provides decentralized applications and advanced security features [1]. It is mainly supported by technologies such as Blockchain Networks (BCNs), smart contracts, Artificial Intelligence (AI), distributed communication protocols and cryptographic techniques. The emergence of Web 3.0 has also boosted new diverse applications. However, there are still several challenges that need to be addressed to guarantee security and trust in decentralized networks. For example, AI needs to be used almost in all emerging Web 3.0 applications [2] and it needs to adapt to privacy preserving value sharing among entities. For security, Public Key Infrastructure (PKI) is used in centralized identity management systems which is expensive and centralized. Moreover, services can be disrupted in case Centralized Authority (CA) makes an error (which can be catastrophic for mission-critical vehicle services). For this reason, Distributed Identity Management (DIM) solutions such as Hyperledger Indy[1] have been proposed to build trust without relying on PKI. By relying on selective disclosure and Zero-Knowledge Proofs (ZKPs), only limited required data elements are provided as proof in DIM solutions. For example, proof that a vehicle user is eligible to participate into Federated Learning (FL) process (e.g., training a model for image recognition) can be validated without providing additional information such as name.

Fig. 1 shows the evolution of digital identity provider with DIM solutions. DIM solutions relies on Decentralized Identifier (DID) and Self-Sovereign Identity (SSI). DID is similar to a username and password pair supported by public/private key pairs. DID is recorded in a BCN by a person, organization or device. Each entity can have multiple DIDs and a different DID to each service so each entity cannot be correlated with multiple services. SSI ensures that each individual can

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.

[1]Online: https://www.hyperledger.org/use/hyperledger-indy, Available: January 2023.

978-3-903176-59-1 ©2023 IFIP

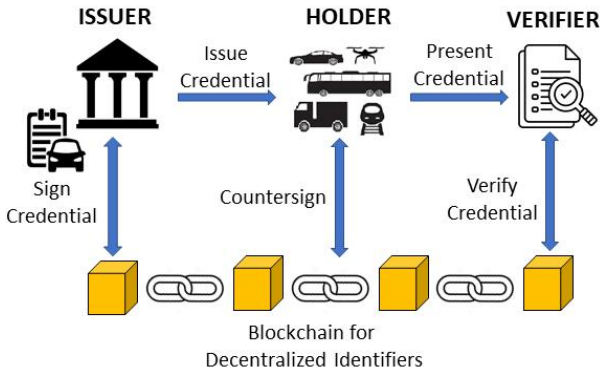


Fig. 1: User centric approach for SSI

gain control of his own data. DID consists of two main parts: (i) unique identifier (produced by owners) and (ii) an associated DID document. DID documents are typically expressed using JSON-LD, containing at least one public key (created by owner), a list of DID authentication ways, services and "verifiable claims". Verifiable claims support two important roles: *(i) claim holders:* an entity that receives and holds claim and *(ii) verifier:* an entity that verifies a claim about a subject.

DIM is still a key open issue in vehicular networks [3] and a solution based on Distributed Ledger Technologys (DLTs) can be very powerful if some of the key issues (e.g., power, network stability, some form of security attack, etc.) can be properly addressed. A BCN-based SSI in a FL architecture for vehicular networks can help overcome some of these limitations. In a BCN-based FL architecture for vehicular networks, some requirements must be met [4]. The first requirement is related to **privacy and security** for user data and preventing unauthorized access or exposure, tampering, malicious updates, or Sybil attacks [5]. The second requirement is **decentralization** of functionalities to avoid a single point of failure and provide scalability [6]. The third requirement is about minimizing the **overhead and computation resource usage**. The fourth requirement is that **trust** should be established between entities involved in the FL process [7].

Some issues for a BCN-based FL architecture in vehicular networks are related to the **limited connectivity** of vehicles in the network to exchange data or participate in the FL process. **Heterogeneous nature** of vehicles in terms of hardware configurations, data types make it difficult to ensure interoperability and data consistency, **dynamic network topology** as vehicles join or leave the FL process, making it challenging to maintain a continuous overview of the network. Some limitations for a BCN-based FL architecture in vehicular networks are **limited bandwidth** for data exchanged during FL, **power constraints** when battery-powered vehicles are involved in FL and **compliance with regulations** related to security and privacy. Some possible solutions are using differential privacy and encryption for privacy, smart contracts and consensus mechanisms for security and integrity of data. A decentralized architecture can be chosen by design. For efficiency, techniques such as compression [8], aggregation [9], or selective sampling [10] can be used to minimize the communication and computational resources required for the FL process when bandwidth constraints and energy saving goals exist. BCN-based mechanisms such as reputation systems and proof-of-work can help build trust between FL participants. Model selection or local aggregation used in edge computing can help with limited connectivity. Normalizing data and approaches such as transfer learning can be used to ensure interoperability and data consistency in the presence of heterogeneity. Finally, consensus mechanisms can help handle dynamic network topologies and enable more efficient data exchange.

II. Related Work and Contributions

There exist many approaches utilizing FL, BCNs, smart contracts, Interplanetary File System (IPFS), SSI ledger, and a Peer-to-Peer (P2P) communication channel for secure, private and efficient data sharing, trust establishment and authentication purposes. In particular, FL has been used with BCNs to provide trust and privacy in several previous works, e.g., in [11] for edge networks, in [6] for applications that integrate FL with BCN for Vehicular Ad Hoc Network (VANET), in [12] for autonomous vehicles, in [13] for Internet of Things (IoT) systems and in [14] for industry 4.0. The research in [4] focuses on the architecture and performance of blockchain-based FL in vehicular networks. A blockchain-based data sharing mechanism for vehicular networks highlighting the importance of data sharing and storage requirements in FL architecture is given in [15]. However, classical FL can suffer from various attacks including adversarial attacks, authentication, and model inversion attacks which need to be enhanced [16]. In the context of identity management and authentication, the authors in [16] provide an identity management and authentication scheme for edge devices utilizing

FL, Veroma[2] provides a platform to create and manage decentralized identifiers and verifiable credentials. At the same time, existing blockchain solutions such as Etherium and Bitcoin are slow in support of DIDs. Hyperledger Indy appears to be a promising solution but is still not widely adapted.

All of the above approaches provide different approaches to the identity management either as a single solution or together with using it with other techniques (such as FL or differential privacy approaches). However, there is a lack of SSI platform for reliable vehicular networks that can provide confidentiality, integrity and authentication all at the same time while keeping the data private (e.g., by relying on techniques such as FL). Our previous works on vehicular networks involved either pure identity management as in [17] or combining it with hierarchical FL methods as in [18]. In this paper we propose a new architecture and design approach to conventional blockchain-based SSI during FL process so that we can ensure confidentiality, authentication, and integrity of vehicle users identity and their data, all simultaneously.

III. Proposed Method

Fig. 2 shows the process of the BCN-based SSI using FL for vehicular networks. In fact, there are two key entities including distributed clients (i.e., vehicular users) and an FL aggregation server. The main idea of the proposed method is to establish trust between clients (or vehicular user in our scenario) and the aggregation server while keeping the identity and data of clients private. The interaction enables secure and transparent execution of ML-related transactions without the need for a central authority. The process consists of six main phases:

Steps (1-2-3) Registration and authentication: When a vehicle user joins to FL system, registration phase is needed to authenticate vehicle user's identity. In this phase, clients will register their IP addresses to the central/aggregation node that is responsible for aggregation of model weights. In case of failure, vehicular user will not be allowed to join the FL protocol.

Step-4) Training: In this phase, once the vehicle users are successfully authenticated, clients train their local models. Most of the time, local training procedure involves running a Stochastic Gradient Descent (SGD) algorithm as shown above.

Step-5) Aggregation: In this phase, aggregation node performs a weighted averaging of the model to generate global model of communication round t. The aggregation rule can be in different forms, such as weighted sum rule or Byzantine fault-tolerant aggregation rules [19], [20].

Step-6) Global Parameters: In this phase, aggregation node sends the global parameters w_{t+1} back to the clients at the iteration $t + 1$. The whole FL procedure starting at step-3 is re-executed until the global model converges.

A. An Example Use Case

One possible use of the proposed FL approach along with BCN-based SSI architecture is in the context of self-driving. Assume that a fleet of autonomous vehicles is in the process of FL (e.g., an image detection or traffic analytics) and aims to improve their Machine Learning (ML) models to generalize and make better decisions. During the initial phase of registration, authenticating vehicles while protecting their privacy in a scalable and reliable manner is a challenging task, due to the need for a central authority that manages and identifies their identities. The FL approach along with BCN-based SSI architecture can assist vehicles in securely and efficiently identifying themselves and sharing FL related information with aggregation server or relevant authorities (e.g., vehicle registration agencies) during the FL process.

B. BCN-based Secure Architecture Enhancements for FL

Fig. 3 shows three methods of network security and privacy for vehicular users relying on SSI-based BCN solution while performing FL. In a classical BCN-based SSI system that uses FL, the verification of the vehicle user's credentials during registration and authentication is done by checking the DID in a permissionless blockchain network as shown in Fig. 3(a). This can be done immediately without the need for the issuer's involvement. Additionally, the vehicle user has the ability to control which attributes are disclosed or kept private during interactions with the FL aggregation server. Fig. 3(b) is an advancement of Fig. 3(a) and

[2]Online: https://veramo.io/, Available: January 2023

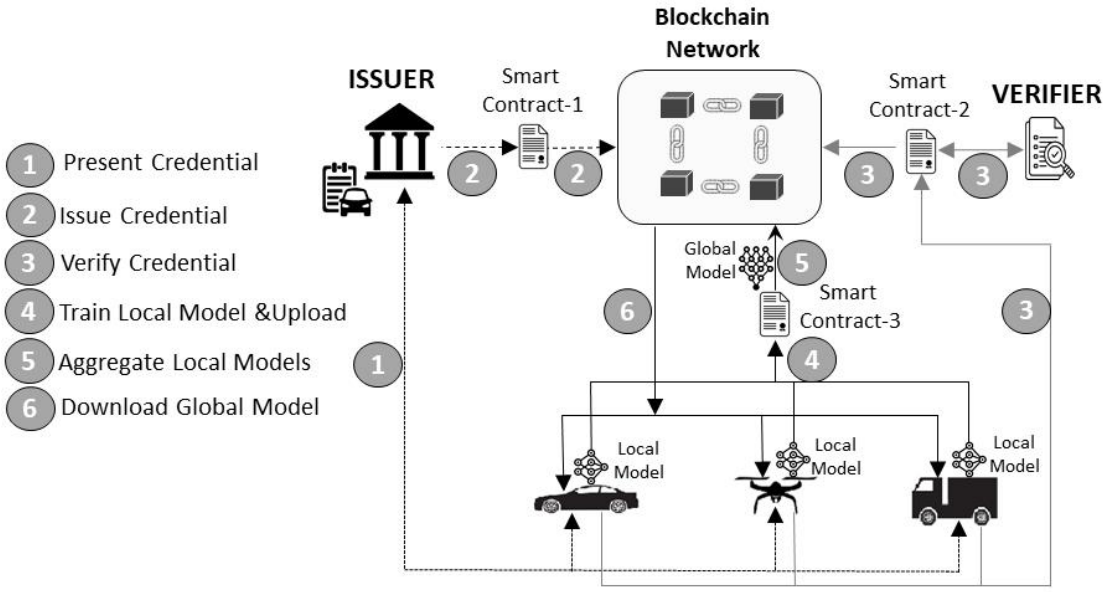


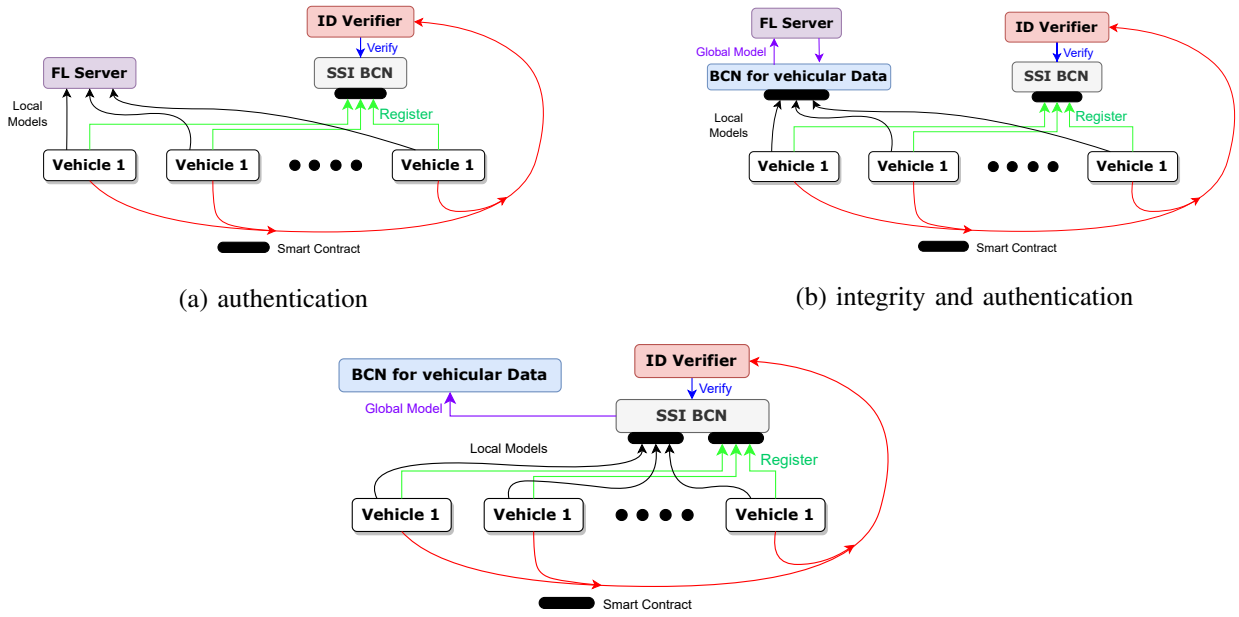
Fig. 2: The integration of decentralized identity concept into a blockchain network for securing the FL-based network operations.

includes the use of another permissionless blockchain network for the permanent and secure recording of other relevant vehicular data. Note that this second BCN provides integrity to each vehicular users's local training model. When vehicle users participate into FL process, they first verify their digital identities by presenting certificates and attestation to an identity verifier that relies on SSI-based BCN. Note that separate transactions at this stage are needed for each user which can be very time-consuming. Once their identities have been verified, the vehicle users can use blockchain network to exchange information with FL aggregation server, increasing the reliability and integrity of the system. Note that each vehicle user also has a transaction account number on the BCN (as vehicular data exchange in FL process (which includes the weights of the model along with other parameters, e.g., Service Level Agreement (SLA) requirements as input to weight update process) are created as transactions), allowing for the matching of their vehicle ID (in the SSI-based BCN) with the account ID number, which can lead to a violation of confidentiality.

Fig. 3(c), on the other hand, aims to ensure the authentication, integrity, and confidentiality of the FL system by using a two-step process. First, vehicle data used during FL process is passed to a permissioned blockchain network for integrity assurance. Next, the permissioned blockchain network requests approval from a permissionless blockchain-based SSI to create blocks for transactions owned by these vehicular users and to verify their validity. If the users are verified in the SSI blockchain, the block is created and the vehicle data is committed to the permissioned blockchain network. With this approach, the vehicle user does not need to perform any authentication process, which ensures confidentiality by preventing the matching of the digital ID with the blockchain account ID of the vehicle user from being disclosed. Additionally, this approach is faster than that of Fig. 3(b) as it eliminates the need for further authentication by the vehicle user in this dynamic and mobile environment.

C. Some Limitations and Possible Solutions

Note that we assumed that the BCN-based SSI registration mechanism would prevent or deter malicious users from participating in the FL process. However, it is possible for a user to join the system with the intention of disrupting the FL process. One possible solution to this problem is to introduce a *reputation system*, where users are assigned a reputation score based on their behavior in the FL process. The reputation score can be based on various factors, such as the accuracy of their models, their contribution to the FL process, and their adherence to the FL protocol. Users with a low reputation score can be excluded from the FL process or given limited access to the FL data. This can help mitigate the risk



(a) authentication

(b) integrity and authentication

(c) confidentiality, integrity and authentication

Fig. 3: Blockchain SSI architectures with FL for vehicular networks using cryptographic techniques and blockchain protocols.

of malicious users disrupting the FL process. Other solutions such as FL with differential privacy (to make it more difficult for malicious users to extract sensitive information or manipulate the FL process by introducing noise into the model updates), FL with with Adversarial Robustness (to make the models more resilient to attacks by malicious users by introducing adversarial examples into the training data) and FL with Multi-Party Computation (to perform secure model aggregation without revealing the models to any party which can help prevent malicious users from manipulating the FL process). Other techniques such as such as secure multi-party computation, homomorphic encryption, and secure enclaves (e.g., to protect the confidentiality and integrity of the model updates) can also be used.

Note also that the usage of BCN would make the discovery of such malicious clients even harder, because (i) it makes any double-checking of the calculations harder to design and execute due to each piece of the FL data being dealt by a different client, and (ii) it potentially increases the computational cost of any such double-checking due to possible additional requirements of the BCN regarding the content stored in the blocks to be checked (in particular with respect to the storage of the relevant portion of the FL state prior to each iteration of the process). On the other hand, the BCN can also provide additional security benefits, such as immutability and transparency. The BCN can be used to record the partial model states during the FL process, which can be used to verify the integrity of the FL process. Additionally, the BCN can provide a transparent audit trail of the FL process, which can help detect any anomalies or suspicious behavior.

The use of a BCN in FL can potentially lead to scalability issues due to the need to reach consensus before moving to the next iteration of the process round. This is because each update to the FL model needs to be validated and recorded on the BCN before the next iteration can begin (as to prevent the waste of computational work due to the formation of orphan nodes), which can slow down the FL process and cause a bottleneck. To address this issue, the FL system can be designed to optimize the consensus process and reduce the impact on scalability. For example, the system can use consensus algorithms that are designed for scalability, such as Proof of Stake (PoS) or Byzantine Fault Tolerance (BFT) consensus algorithms. These algorithms can improve the scalability of the FL process by reducing the computational overhead and minimizing the number of messages exchanged between nodes.

The behavior of the proposal approach in the situation of increasing complexity and size of local weights in FL can be evaluated based on several factors. While storing the local weights of client models on the BCN ledger provides data protection and security benefits, it may pose challenges when it comes to scalability and performance. When the size of the local weights increases, storing them directly on the BCN ledger can lead to scalability issues. The BCN has inherent limitations in terms of storage capacity and transaction throughput. Storing large amounts of data on the BCN can result in increased storage requirements and slower transaction processing times, which can impact the overall performance of the system. To address this, alternative approaches can be considered. One option is to store only a file validator (CRC) in the BCN ledger instead of the entire weight data. This approach reduces the storage requirements on the BCN and improves scalability. The weights can be transferred directly between the client models and the aggregator without being stored on the BCN. The file validator can serve as a proof of integrity, ensuring that the weights have not been tampered with during the transfer process. The suitability of each approach, whether storing the weights directly on the BCN or using a file validator, depends on the specific requirements of the BCN network and the use case. It is important to evaluate the trade-offs between security, scalability, performance, and storage requirements within the context of the proposed architecture.

IV. Simulations

In our testing environment, we utilize the OpenStack platform and create two virtual machines. One virtual machine hosts a Representational State Transfer (REST) server responsible for generating requests to Hyperledger Indy, while the other virtual machine is configured with a container-based setup for the DID system. Hyperledger Indy serves as the blockchain for managing DIDs, acting as the identity layer for our test setup. The DID virtual machine is allocated 8 GB RAM, 4 vCPUs, and 40 GB disk space within OpenStack. During the installation of Hyperledger Indy, we set up the VON network[3], which enables the operation of container-based nodes. Specifically, four DID container nodes are created to serve as DID issuers. The REST API, implemented using the Flask framework, facilitates user requests for the BCN. To communicate with the BCN, we employ the DIDComm messaging protocol, an encrypted communication protocol developed as part of the Hyperledger Aries project[4].

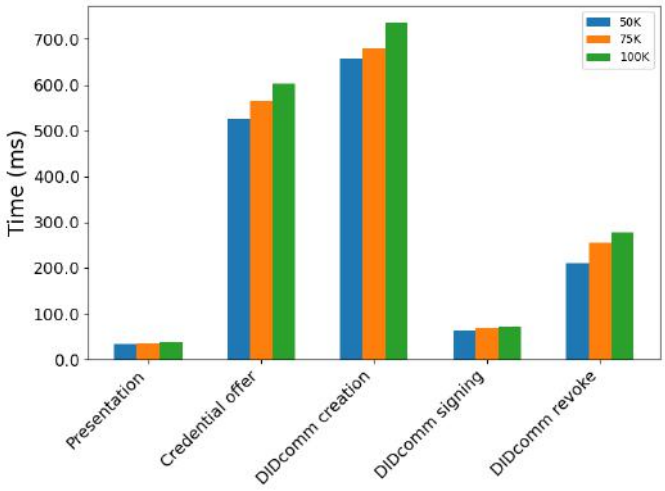


Fig. 4: Average values of experimental DID implementation for credential operations.

Based on the experimental results presented in Fig. 4, the following observations can be made regarding the performance of the credential operations: *The average credential presentation time,* which measures the time taken to present credentials to the verifier, ranges from 33 milliseconds (for 50,000 (50K) requests) to 39 milliseconds (for 100,000 (100K) requests). There is a slight increase in presentation time as the number of requests increases. *The average credential offer time,* which quantifies the time to generate and offer credentials to the holder, ranges from 525 milliseconds (for 5K requests) to 603 milliseconds (for 100K requests). Similar to presentation time, there is a slight increase with an increasing number of requests. *The average DIDcomm connection creation time,* evaluating the time to establish a connection using the DIDComm protocol, ranges from 659 milliseconds (for 50K requests) to 736 milliseconds (for 100K requests). There is a moderate increase in connection creation time as the number of requests increases. *The average DIDcomm signing time,* measuring the time to sign messages using the DIDComm protocol, ranges from 64 milliseconds (for 50K requests) to 72 milliseconds (for 100K requests). The signing time

[3]VON Network, Available: https://github.com/bcgov/von-network, Online: March 2023.

[4]Aries RFC 0005: DID Communication, Available: https://bit.ly/3TKFUKG, Online: March 2023.

remains relatively consistent across different numbers of requests. *The average DIDcomm revoke credential time,* indicating the time to revoke a credential, ranges from 212 milliseconds (for 50K requests) to 277 milliseconds (for 100K requests). Similar to other metrics, there is a slight increase in revoke credential time as the number of requests increases. These results provide insights into the performance characteristics of the credential operations in the implemented system. Most operations demonstrate efficient execution with average times within reasonable ranges. However, it is worth noting that there is a slight degradation in performance as the number of requests increases, particularly for offering credentials and creating DIDcomm connections.

V. Conclusion

DIM and FL fields offer a diverse array of opportunities paving the way for getting control over the usage of shared and not shared vehicular data with third parties. In this paper, we considered three different architectural options to provide either only authentication or both authentication and integrity or confidentiality, authentication and integrity to FL process. We also provide an example use case considering the proposed BCN-based SSI solution for FL of autonomous networks. The proposed architectures are expected to provide a secure and reliable platform for sharing the data. Experimental results in credential operations indicate there is a slight increase in presentation time, offer time, connection creation time, and revoke credential time as the number of requests increases, suggesting a slight degradation in performance for these operations.

References

[1] C. Chen *et al.*, "When digital economy meets web 3.0: Applications and challenges," *IEEE Open Journal of the Computer Society*, 2022.

[2] T. Issa, *Artificial intelligence technologies and the evolution of web 3.0*. IGI Global, 2015.

[3] P. Fraga-Lamas and T. M. Fernández-Caramés, "A review on blockchain technologies for an advanced and cyber-resilient automotive industry," *IEEE access*, vol. 7, pp. 17578–17598, 2019.

[4] N. Sehar, O. Khalid, I. A. Khan, F. Rehman, M. A. Fayyaz, A. R. Ansari, and R. Nawaz, "Blockchain enabled data security in vehicular networks," *Scientific Reports*, vol. 13, no. 1, p. 4412, 2023.

[5] J. Grover, "Security of vehicular ad hoc networks using blockchain: A comprehensive review," *Vehicular Communications*, p. 100458, 2022.

[6] A. R. Javed *et al.*, "Integration of blockchain technology and federated learning in vehicular (IoT) networks: A comprehensive survey," *Sensors*, vol. 22, no. 12, p. 4394, 2022.

[7] S. Otoum, I. Al Ridhawi, and H. T. Mouftah, "Blockchain-supported federated learning for trustworthy vehicular networks," in *GLOBECOM 2020-2020 IEEE Global Communications Conference*, pp. 1–6, IEEE, 2020.

[8] S. M. Shah and V. K. Lau, "Model compression for communication efficient federated learning," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[9] L. Liu, J. Zhang, S. Song, and K. B. Letaief, "Client-edge-cloud hierarchical federated learning," in *ICC 2020-2020 IEEE International Conference on Communications (ICC)*, pp. 1–6, IEEE, 2020.

[10] 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.

[11] D. C. Nguyen *et al.*, "Federated learning meets blockchain in edge computing: Opportunities and challenges," *IEEE Internet of Things Journal*, vol. 8, no. 16, pp. 12806–12825, 2021.

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

[13] M. Ali *et al.*, "Integration of blockchain and federated learning for internet of things: Recent advances and future challenges," *Computers & Security*, vol. 108, p. 102355, 2021.

[14] Y. Qu *et al.*, "A blockchained federated learning framework for cognitive computing in industry 4.0 networks," *IEEE Trans. on Industrial Informatics*, vol. 17, no. 4, pp. 2964–2973, 2020.

[15] M. Rehman, Z. A. Khan, M. U. Javed, M. Z. Iftikhar, U. Majeed, I. Bux, and N. Javaid, "A blockchain based distributed vehicular network architecture for smart cities," in *Web, Artificial Intelligence and Network Applications: Proceedings of the Workshops of the 34th International Conference on Advanced Information Networking and Applications (WAINA-2020)*, pp. 320–331, Springer, 2020.

[16] R. U. Haque *et al.*, "Towards convergence of blockchain and self-sovereign identity for privacy-preserving secure federated learning," in *Big Data and Security: Third International Conference*, pp. 243–255, Springer, 2022.

[17] E. Zeydan, J. Mangues, S. Arslan, and Y. Turk, "Blockchain-based self-sovereign identity solution for vehicular networks," in *2023 19th International Conference on the Design of Reliable Communication Networks (DRCN)*, pp. 1–7, IEEE, 2023.

[18] E. Zeydan, J. Mangues, S. S. Arslan, and Y. Turk, "Self-sovereign identity management for hierarchical federated learning in vehicular networks," in *2023 IEEE 24th International Conference on High Performance Switching and Routing (HPSR)*, pp. 191–196, IEEE, 2023.

[19] S. Li, E. Ngai, and T. Voigt, "Byzantine-robust aggregation in federated learning empowered industrial iot," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 2, pp. 1165–1175, 2021.

[20] J. Yang *et al.*, "Clean-label poisoning attacks on federated learning for IoT," *Expert Systems*, p. e13161, 2022.