Towards FAIR Data in Distributed Machine Learning Systems

Yongli Mou[1]*, Fengyang Guo[2]*, Wei Lu[1], Yongzhao Li[1],
Oya Beyan[3,4], Thomas Rose[1,4], Schahram Dustdar[2], and Stefan Decker[1,4]
[1]*RWTH Aachen University, Aachen, Germany*
[2]*Distributed Systems Group, TU Wien, Vienna, Austria*
[3]*Institute for Biomedical Informatics, University of Cologne,*
Faculty of Medicine and University Hospital Cologne, Cologne, Germany
[4]*Fraunhofer FIT, Sankt Augustin, Germany*

GLOBECOM 2023 - 2023 IEEE Global Communications Conference | 979-8-3503-1090-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/GLOBECOM54140.2023.10437414

***Abstract*—In the era of big data and artificial intelligence, distributed machine learning has emerged as a promising solution to address privacy and security concerns while fostering collaboration between multiple parties. However, with the data increased in terms of volume, velocity, veracity and variety, ensuring effective data management and responsible data sharing in these systems remains a challenge. In this paper, we explore the potential solutions and propose a system architecture that incorporates FAIR data principles (Findable, Accessible, Interoperable, and Reusable) to promote effective and secure collaboration in federated learning. A minimum set of metadata schemes tailored for distributed machine learning and a decentralized authentication and authorization mechanism based on self-sovereign identity and policy-based access control architecture are proposed. To demonstrate the effectiveness of the proposed system, we conduct a FAIRness assessment and evaluate the model performance with a federated learning use case. Our work contributes to the development of an efficient, secure, and collaborative data ecosystem, fostering innovation in artificial intelligence and machine learning.**

***Index Terms*—FAIR data principles, distributed machine learning, federated learning**

I. Introduction

With the rapid development of the Internet of Things (IoT), vast amounts of data are generated ever-increasingly in terms of volume, variety, velocity, and veracity [1]. Data has become an indispensable asset for individuals, businesses, organizations and governments, thanks to the rapid development of sophisticated analytical tools and artificial intelligence techniques such as deep learning, providing valuable insights to improve business operations and support policy decisions. Nonetheless, privacy and security concerns raise barriers to data collection and data sharing. In addition, the prevalence of isolated data islands in various industries due to factors such as competitive rivalry and complex administrative processes resulting in a fragmented data ecosystem has impeded data-driven innovation [2]. Distributed machine learning has emerged as a promising response to these challenges, enabling multiple parties to cooperate and train machine learning models without disclosing their raw data by shifting the paradigm from data centralization to bringing computation to the data.

This paradigm shift towards distributed machine learning has underscored the need for effective and responsible data management and stewardship that facilitates seamless cooperation and responsible data sharing among the participating parties. The FAIR data principles [3] have emerged as guiding principles that provide a set of guidelines to improve the *Findability*, *Accessibility*, *Interoperability* and *Reuse* of research data. The FAIR principles apply to digital resources such as datasets, code, workflows, or other research objects, as well as metadata and supporting infrastructure (e.g., search engines) [4]. By adhering to the FAIR principles, stakeholders in distributed machine learning systems can ensure that the data utilized in these systems is discoverable, efficiently and securely managed, thus maximizing the value derived from the data while minimizing potential risks.

In this paper, we explore the challenges and potential solutions for incorporating FAIR data principles in distributed machine learning systems. We propose a novel method to further enhance FAIR data compliance in distributed machine learning systems. Our contributions are as follows:

- We leverage Decentralized Identifiers (DIDs), which are decentralized, cryptographically verifiable, globally unique, and resolvable, as Persistent Identifiers (PIDs) to ensure the findability of digital objects.
- We develop a minimum set of metadata schemes tailored for distributed machine learning, which enhances data discoverability and interoperability.
- We propose a novel system architecture that consists of a data space for metadata services, and the blockchain network and supports secure and stable collaboration among data providers and data consumers.
- We utilize DIDs and Verifiable Credentials (VCs) for Self-Sovereign Identity (SSI) and based on it we develop decentralized policy-based authentication and authorization architecture to secure data access and control in distributed machine learning systems.
- Furthermore, we demonstrate the effectiveness of the proposed solutions by the evaluation of the FAIRness of the implemented prototype system and of the model

*These authors contributed equally to this work.

performance with the federated learning use cases.

II. Background

In this section, we provide an overview of the key concepts underlying our work, namely the FAIR data principles, federated learning, and distributed ledger technologies along with the smart contract.

A. FAIR Data Principles

The FAIR data principles are guiding principles for scientific data management and stewardship that provide a set of guidelines initially proposed to improve the *Findability*, *Accessibility*, *Interoperability* and *Reuse* of the research data [3]. Most of the requirements for findability and accessibility can be achieved at the metadata level. Interoperability and reuse require more effort at the data level. The key points of the FAIR data principles highlight the following specific components and features:

- Persistent identifiers: Unique and stable identifiers assigned to data, ensuring long-term findability and accessibility.
- Rich metadata: Detailed descriptions of data, making it easier to discover, understand, and use by both humans and machines.
- Standardized formats and vocabularies: Consistent use of data formats, terminologies, and ontologies to enhance interoperability and facilitate data exchange.
- Open protocols: Adoption of open standards and protocols for data access, ensuring transparency and fostering collaboration among different stakeholders.
- Clear licensing and usage policies: Well-defined policies guiding data reuse, ensuring proper attribution, and compliance with legal and ethical guidelines.

B. Federated Learning

Federated learning [5] is a distributed privacy-preserving machine learning paradigm that enables multiple parties to jointly train machine learning models without disclosing their raw data, and involves training a global model by aggregating local model updates from multiple data sources. A system for federated learning consists typically of a server and a collection of clients.

Given the global dataset D distributed on K clients, where each client $k \in \{1, 2, \ldots, K\}$ holds a local dataset D_k of size n_k, we denote the global dataset as the union of all local datasets, i.e., $D = \bigcup_{k=1}^{K} D_k$, and has a total size of $n = \sum_{k=1}^{K} n_k$.

In each round of federated learning, each client k computes an update to the model parameters based on its local dataset D_k. Let the model parameters be denoted by $\boldsymbol{w}$, and the objective function for the optimization problem be $L(\boldsymbol{w})$. The local objective function for client k can be defined as in Equation 1, where $l(\boldsymbol{w}; x_i, y_i)$ is the loss function for a single data point (x_i, y_i).

$$L_k(\boldsymbol{w}) = \frac{1}{n_k} \sum_{i \in D_k} l(\boldsymbol{w}; x_i, y_i) \tag{1}$$

The global objective function, which we aim to optimize, is the weighted average of the local objective functions, as shown in Equation 2.

$$L(\boldsymbol{w}) = \sum_{k=1}^{K} \frac{n_k}{n} L_k(\boldsymbol{w}) \tag{2}$$

In each round, client k computes a local update $\Delta \boldsymbol{w}_k$ by minimizing the local objective function $L_k(\boldsymbol{w})$. The server then aggregates the local updates from all clients to compute the global update, which is given by Equation 3.

$$\Delta \boldsymbol{w}_{\text{global}} = \sum_{k=1}^{K} \frac{n_k}{n} \Delta \boldsymbol{w}_k \tag{3}$$

Finally, the server updates the global model parameters with the global update using Equation 4, where η is the learning rate and t denotes the current round of federated learning.

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t + \eta \Delta \boldsymbol{w}_{\text{global}} \tag{4}$$

This process continues and repeats for a predefined number of rounds or until a convergence criterion is met.

One of the challenges in federated learning is dealing with non-Independent and Identically Distributed (non-IID) data. Non-IID data arises when the data distribution across different parties varies, which is often the case in real-world scenarios and can lead to reduced model performance and slow convergence rates, as the global model may struggle to generalize well across heterogeneous data sources [6].

C. Blockchain

Distributed ledger technologies (DLTs), such as blockchain, are decentralized systems that enable secure and transparent data storage and management, in which transactions are stored in a series of blocks connected with each other and the latter block records the hash value of the former block. Once a block has been confirmed, the transactions in that block can not be changed. They use cryptographic techniques and consensus algorithms to ensure data immutability, traceability, and tamper resistance [7]. DLTs have the potential to transform various industries by enhancing trust, security, and efficiency in data exchange and collaboration.

One of the key features of DLTs is the ability to create and deploy smart contracts. Smart contracts are self-executing programs with the terms of the agreement between interested parties that are deployed on the blockchain and can automatically enforce predefined rules and conditions encoded within them. They can facilitate, verify, and enforce the performance of a contract without the need for intermediaries. The output of a smart contract is stored in a transaction. Same as the other transactions on the blockchain, these transactions are traceable and irreversible [8]. In the context of distributed machine learning systems, smart contracts can be used to automate and enforce data-sharing agreements, model update processes, and access control mechanisms, thereby enhancing security, transparency, and efficiency.

III. Proposed System

In this section, we describe the proposed distributed learning system designed to incorporate FAIR data principles into federated learning. We begin with a high-level overview of the system architecture and then delve into the details of the metadata for distributed machine learning. Next, we illustrate the decentralized authentication and authorization process between different entities in the system. Then, we describe the model verification for robust aggregation during the federated learning process. Finally, we develop a federated learning prototype of our proposed system.

A. System Overview

Our proposed system aims to facilitate efficient and secure collaboration in federated learning by implementing the FAIR data principles. Figure 1 gives an overview of the system architecture that enables the collaboration of data owners and data consumers for federated learning, which includes the following key components:

- Data Space (DS): An underlying metadata infrastructure providing APIs for metadata services where users can register, store and manage the metadata of their distributed machine-learning digital assets, ensuring that data is findable, accessible, and interoperable.
- Blockchain Network: A distributed infrastructure with smart contracts enables the secure and transparent management of the metadata, federated learning processes, and access control.
- Data Consumers (DCs): Stakeholders with the federated learning aggregation server infrastructure, which is responsible for aggregating model updates from multiple Data Providers (DPs).
- Data Providers (DPs): Stakeholders who own the infrastructure for federated learning clients, to which data is attached. DPs are responsible for managing the data collection, training local models on their data and sharing model updates with the federated learning server.

Both DCs and DPs have digital agents that facilitate interactions with the blockchain network to securely and privately manage various types of data, such as personal dataset metadata, algorithm metadata, and policy definitions for datasets and algorithms.

In the proposed system, federated learning algorithms are divided into the client algorithm for training and verification, which is owned by DPs and the server algorithm for aggregation, which is owned by DCs. These algorithms are encapsulated in an Open Container Initiative-compliant[1] image (such as a Docker image) separately. By taking advantage of containerization technologies, the execution of federated learning is an operating system (OS)-independent, since the necessary dependencies to run the code such as libraries and packages are captured within the image. Consequentially, the algorithm can be written in any programming language, such as Python or R. Beyond these advantages, our proposed system

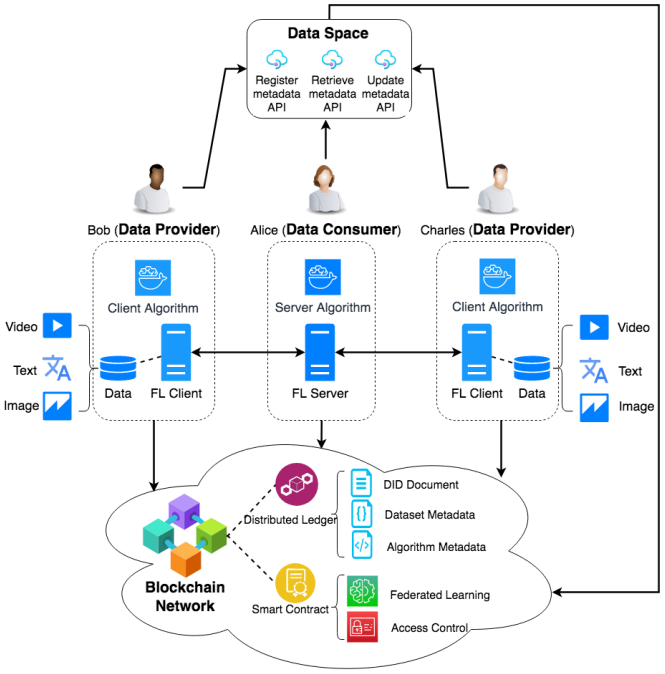

Fig. 1: System Architecture

is data source agnostic, i.e., the choice of data sources is flexible since the data are attached during the execution.

By integrating these components, the proposed system addresses various requirements related to FAIR principles in federated learning, paving the way for a more efficient secure, and collaborative data ecosystem for DCs and DPs.

B. Metadata for Distributed Machine Learning

In order to implement FAIR data principles and ensure the findability, accessibility, and interoperability of the metadata, we develop Persistent Identifiers (PIDs) and a minimal metadata scheme for our proposed distributed machine learning system.

PIDs are fundamental to FAIR data ensuring that data can be reliably cited and referenced, required to be globally unique, resolvable, and associated with a set of additional metadata [9]. DIDs are a kind of self-sovereign identifier registered on a distributed ledger that provides a globally unique digital identity for any entity or object. We propose to use DIDs as PIDs because they are not only globally unique required by FAIR principles, but also decentralized, self-sovereign, and interoperable.

Rich metadata and standards, such as vocabularies and ontologies, are essential components of FAIR data and good drivers for enhancing findability as well as interoperability and reuse. The concept of the FAIR Data Point (FDP) is developed as an exemplary lightweight infrastructure component and demonstrated compliance with the FAIR guiding principles [10]. FDP uses global standards (e.g., Dublin Core[2]

[1]https://opencontainers.org/

[2]https://www.dublincore.org/specifications/dublin-core/

and DCAT[3]) to provide access to structured metadata. Our proposed FAIR metadata schema is based on FDP Metadata and DCAT metadata with basic fifteen attributes to describe the basic information of datasets and algorithms.

Additionally, we adopt the metadata schema in [11] for distributed machine learning datasets and algorithms. For datasets, we add three additional attributes: purposes, gaps, and tasks. Purposes refer to the intended usage of the dataset in relation to the algorithm type. Gaps define the specific areas the dataset aims to address. Tasks describe the specific machine learning tasks for which the dataset is designed. Similarly, we add three attributes to the algorithm metadata: version, entry point, and parameters. The version attribute represents the current iteration of the algorithm. The entry point specifies for example the URL of the algorithm's Docker image. Parameters encompass the necessary input parameters for the algorithm to function correctly. Listing 1 and 2 show simple examples of the metadata of CIFAR-10 dataset and FedAvg algorithm in JSON.

```
{
    ...
    "dmls:Identifier": "did:dmls:91d7ac8d-ac
    f0-41cb-a256-18101f22dd04"
    "dmls:Publisher": "did:dmls:f4911c39-ae
    b8-468d-bf43-02c84c05af8a"
    "dmls:Title": "The CIFAR-10 dataset"
    "dmls:Purposes": "Image classification"
    "dmls:Gaps": "Availability of labeled
    image datasets for machine learning and
    computer vision research"
    "dmls:Tasks": "Image classification"
}
```

Listing 1: Metadata Example of the CIFAR-10 Dataset

```
{
    ...
    "dmls:Identifier": "did:dmls:91d7ac8d-
    acf0-41cb-a256-18101f22adbf"
    "dmls:Publisher": "did:dmls:f4911c39-ae
    b8-468d-bf43-02c84c05ajdhc"
    "dmls:Title": "Federated Averaging
    algorithm"
    "dmls:Version": "v0.0.1"
    "dmls:Entrypoint": "localhost:8080"
    "dmls:Parameters": "Batch size, learning
    rate, number of communication rounds,
    client selection strategy, regularization"
}
```

Listing 2: Metadata Example of the FedAvg Algorithm

As a result, the metadata schema of the dataset and the algorithm describes all the necessary resource information. Both Dcs and DPs in the system are able to retrieve and serialize JSON files. Once all machine learning objects are registered according to the proposed metadata schema, the DC and DP create access control policies for secure interaction.

C. Decentralized Authentication and Authorization

In our proposed system, we achieve decentralized authentication and authorization with Self-Sovereign Identity (SSI) by combining blockchain-empowered DIDs and VCs in the access control architecture.

[3]https://www.w3.org/TR/vocab-dcat-2/

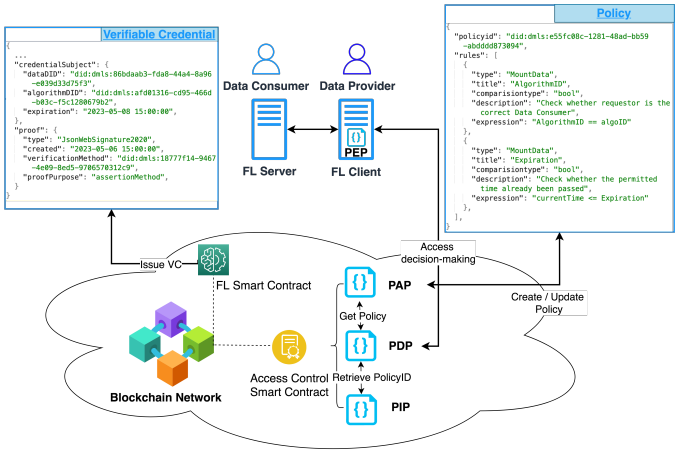

Fig. 2: Access Control Architecture

The authentication process is in the following steps:

1. The DC and DP register in the system through the blockchain smart contract and are assigned DIDs that are stored in the blockchain network.
2. The DC searches for desired datasets using the APIs provided by the Data Space and sends an authentication request to the corresponding DPs.
3. Upon receiving the request, the DPs verify the DC's DID using the evidence provided in the DID document.

If the DP grants the DC permission to use the data, the DC receives a VC issued by the DPs through the FL contract. At the same time, the DP makes the specific training datasets available on the FL client.

Figure 2 shows the access control architecture for the authorization process, which is built upon SSI and includes four components: the Policy Information Point (PIP), Policy Administration Point (PAP), Policy Decision Point (PDP), and Policy Enforcement Point (PEP). The first three components are smart contracts deployed on the blockchain. The PIP provides the policy ID for the requested resource of DC. The PAP handles policy creation and updates. The PDP makes access decisions. The PEP is a bridge between the FL client and PDP, and integrates the blockchain network gateway component and interacts with PDP's exposed APIs. The authorization process is described below:

1. DCs and DPs create policies on personal resources through the PAP smart contract and store them locally.
2. FL server sends the access requests to the FL Client. The request contains associated DIDs of the dataset and algorithm.
3. The PEP reads the locally stored VC and sends the required attributes to the PDP for decision-making.
4. The PDP smart contract calls the PIP to get the policy DID, which is created on the requested resource.

5. The PDP smart contract calls the PAP smart contract to get the local policy through the policy DID.
6. The decision-making function in the PDP smart contract evaluates the rules with the credential subject to make a decision, which is then sent back to the PEP.

The PEP extracts a boolean variable as the operational criterion. If true, the federated learning client can initiate training; otherwise, training is disallowed. The decision is stored in the blockchain network ledger and can be queried for verification purposes.

D. Robust Aggregation with Model Verification

In order to achieve a robust aggregated global model, we develop a model verification process before the aggregation. Initially, the aggregation server at the DC divides all local clients at DPs into training clients and verification clients.

Then the training clients begin model training on their own datasets. After each training round, training clients submit their model updates and local model weights to the aggregation server. Next, the aggregation server sends the model updates to verification clients for model assessment. Verification clients test the updates on mixed datasets and return various metrics, such as accuracy, error rate, precision, recall, and F1 score, based on testing results. The aggregation server evaluates these metrics and decides whether to aggregate the model updates or not. If the results meet aggregation requirements, the server proceeds with aggregation and submits the training record and the qualified model update hash to the blockchain. Otherwise, it discards the model updates and initiates a new training round.

Compared to the traditional federated learning process, incorporating model verification offers several advantages. By introducing a verification step, the system ensures that model updates from various nodes contribute positively to the global model's performance. This helps maintain high-quality models and prevents poor updates from impacting the overall learning process and potentially address the non-IID problem.

E. Blockchain-empowered Federated Learning

We develop a prototype of our proposed distributed machine learning system with the blockchain-empowered process.

The motivation for incorporating blockchain into federated learning is three-fold: audibility, provenance and trust. We develop smart contracts for the federated learning process that automate and enforce agreements between parties, ensuring adherence to the agreed-upon protocols. Blockchain's immutable nature provides an ideal solution for maintaining an auditable record of model updates and training history. This enables tracking the model's provenance. Blockchain establishes trust among these nodes by offering a transparent and verifiable record of transactions.

IV. Evaluation

A. FAIRness Analysis

To evaluate the FAIRness of the proposed system, we analyze the corresponding properties of different components in the system. The following analysis with **F**, **A**, **I** and **R** is based on the definition of the FAIR principles [4].

F1: Each dataset metadata and algorithm metadata is assigned a globally unique and persistent DID. Users can locate the specific target dataset or algorithm by using the globally unique DID. As such, the system satisfies the F1 requirement.

F2: All datasets and algorithms are described with extensive metadata vocabularies, which encompasses all the necessary attributes to fully describe a dataset or algorithm in a machine learning system. Therefore, the system satisfies the F2 requirement.

F3: The dataset and algorithm metadata include the publisher DID of the resource, which can be used to locate the original resource in the publisher's database. So the system fulfills the F3 requirement.

F4: The metadata is stored in the Fabric sample network ledger indexed by its globally unique DID, which is obviously a searchable repository. So the system fulfills the F4 requirement.

A1.1: The blockchain network can be read and written through a gateway using a free and open communication protocol, enabling efficient and standardized access to the metadata. For example, the Ethereum Gateway is implemented by JSON-RPC protocol and the Hyperledger Fabric Gateway uses gRPC protocol. Thus, the system satisfies the A1.1 requirement.

A1.2: The blockchain gateway protocol needs a registered user identity certificates and signature to authenticate and authorize the gateway client access permission. Thus, the system fulfills the A1.2 requirement.

A2: The metadata is accessible as long as the blockchain network is deployed and the key-value pair is not deleted. It does not depend on whether the original resource is available. Even if the publisher does not share the usage of their dataset or algorithm, the dataset or algorithm metadata is still accessible to all users in the blockchain network. Therefore, the system fulfills the A2 requirement.

I1: The metadata is based on FDP Metadata and DCAT Metadata, represented in JSON format, which is machine-readable and can be easily processed by various software programs. Additionally, the Client Algorithm created by the DP, which is containerized, is also interoperable. Therefore, the system fulfills the I1 requirement.

I2: The metadata schema currently incorporates standardised vocabularies that provide the necessary terms or concepts to represent their content. Simultaneously, the vocabularies are retrievable using the globally unique identifier. Consequently, the system fully satisfies the I2 criterion of the FAIR principles.

I3: The metadata includes a qualified reference to other data. So the system fulfills the I3 requirement.

R1.1: The Decentralized Authentication and Authorization functionality offers strictly limited usage rights of the datasets and algorithms, which fulfills the R1.1 requirement.

[4]https://www.go-fair.org/fair-principles/

R1.2: Each dataset and algorithms metadata contains the ”dmls:Relation” attribute, which specifies where the resource is from. The ”dmls:Creator” indicates who generates and collects it. Thus, the requirement R1.2 is satisfied.

R1.3: All datasets and algorithms metadata are implemented in a standardized way and include common vocabularies. Thus, the requirement R1.3 is fulfilled.

In conclusion, the proposed system satisfies most aspects of the FAIR Principles, demonstrating that the whole system is findable, accessible, interoperable, and reusable. Additionally, the metadata is distributed, visible, transparent, accessible, and tamper-proof, keeping the characteristics of the blockchain technology intact.

B. Model Performance

To demonstrate the effectiveness of our proposed federated learning aggregation with the model verification process, we conduct experiments for 5 runs with 10 clients on the CIFAR-10 dataset, which consists of 50,000 training samples and 10,000 test samples and is distributed evenly across 10 distinct classes. Each client holds a subset of the CIFAR-10 dataset and trains a Convolutional Neural Network (CNN) on their local data for 50 rounds. After each round of aggregation, the global model is evaluated on the test dataset deployed on the server. Accuracy and loss are reported as the evaluation metrics for the model performance.

Our experimental results are shown in Figure 3. The difference between the lower and upper bounds of the loss values and accuracy values is generally smaller for the system with verification, which implies that the model with verification is more consistent. The average values of accuracy for the system with verification are consistently higher or comparable to those without verification across all rounds, while the average loss values are lower. When comparing the loss values between the two systems, we can observe that the system with verification tends to converge faster.

In summary, the integration of the verification process in our proposed model aggregation method for federated learning helps maintain high-quality models by filtering out poor updates, thus resulting in improved model performance, faster convergence, and increased consistency across clients.

V. Conclusion

In this paper, we present a novel system that aims to incorporate FAIR data principles into federated learning ecosystems. Our proposed system integrates key components such as metadata for datasets and algorithms, decentralized authentication and authorization mechanisms for secure data access control, and robust aggregation with model verification. Our system fosters a more efficient and collaborative data ecosystem in federated learning, adhering to the FAIR data principles and promoting data-driven innovation across various industries. As future work, we plan to further develop and evaluate the performance of our proposed system in real-world federated learning scenarios, while continuing to explore novel techniques and approaches to address the challenges of non-IID data in distributed machine learning systems.

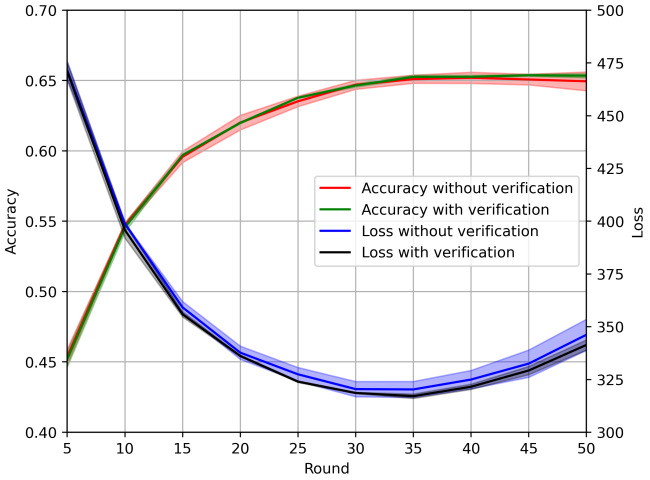

Fig. 3: Model Performance Evaluated on CIFAR-10 Test Set

Acknowledgment

This work was supported by the German Research Foundation DFG project NFDI4Health (grant no. 442326535) and by the German Ministry for Research and Education BMBF project WestAI (grant no. 01IS22094D).

References

[1] S. Sagiroglu and D. Sinanc, “Big data: A review,” in *2013 international conference on collaboration technologies and systems (CTS)*. IEEE, 2013, pp. 42–47.

[2] Q. Yang, Y. Liu, T. Chen, and Y. Tong, “Federated machine learning: Concept and applications,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–19, 2019.

[3] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne *et al.*, “The fair guiding principles for scientific data management and stewardship,” *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[4] M. Thompson, K. Burger, R. Kaliyaperumal, M. Roos, and L. O. B. da Silva Santos, “Making fair easy with fair tools: From creolization to convergence,” *Data Intelligence*, vol. 2, no. 1-2, pp. 87–95, 2020.

[5] 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.

[6] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, “On the convergence of fedavg on non-iid data,” *arXiv preprint arXiv:1907.02189*, 2019.

[7] Z. Zheng, S. Xie, H.-N. Dai, X. Chen, and H. Wang, “Blockchain challenges and opportunities: A survey,” *International journal of web and grid services*, vol. 14, no. 4, pp. 352–375, 2018.

[8] S. Wang, Y. Yuan, X. Wang, J. Li, R. Qin, and F.-Y. Wang, “An overview of smart contract: architecture, applications, and future trends,” in *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2018, pp. 108–113.

[9] N. Juty, S. M. Wimalaratne, S. Soiland-Reyes, J. Kunze, C. A. Goble, and T. Clark, “Unique, persistent, resolvable: Identifiers as the foundation of fair,” *Data Intelligence*, vol. 2, no. 1-2, pp. 30–39, 2020.

[10] L. O. B. da Silva Santos, K. Burger, R. Kaliyaperumal, and M. D. Wilkinson, “Fair data point: A fair-oriented approach for metadata publication,” *Data Intelligence*, pp. 1–21, 2022.

[11] J. Giner-Miguelez, A. Gómez, and J. Cabot, “A domain-specific language for describing machine learning dataset,” *arXiv preprint arXiv:2207.02848*, 2022.