A Data Sharing Privacy Protection Model Based on Federated Learning and Blockchain Technology

Fei Ren*, Zhi Liang
Department of Public Technical Service, State Information Center, Beijing, 100045, China

***Abstract*—As the main driving force for social development in the new era, data sharing is controversial in terms of privacy and security. Traditional privacy protection methods are a bit challenging when faced with complex and massive shared data. Given this, firstly, the Byzantine consensus algorithm in blockchain technology was elaborated. Meanwhile, a decision tree algorithm was introduced for node classification optimization, and a new consensus algorithm was proposed. In addition, local data training and updating were achieved through federated learning, and a new data-sharing privacy protection model was proposed after jointly optimizing consensus algorithms. The maximum throughput of the optimized consensus algorithm was 1560. The maximum consensus delay was 110 milliseconds. After multiple iterations, the removal rate of the Byzantine nodes reached 56.6%. The optimal reputation value of the new data-sharing privacy protection model was 0.75. The lowest reputation value after 10 iterations was 0.32. As a result, this proposed model achieves excellent results in data sharing privacy protection tasks, demonstrating high model feasibility and effectiveness. The research aims to provide a reliable method for data sharing privacy protection in the field.**

***Keywords*—Federated learning; blockchain; data sharing; privacy; reputation**

I. Introduction

The rapidly developing information technology has fully utilized data sharing in various fields such as education, healthcare, and manufacturing [1]. At present, privacy protection has become a major challenge faced by data sharing. Traditional centralized data sharing models carry the risk of privacy breaches, especially when dealing with complex and large amounts of shared data. To address this issue, many researchers have proposed measures such as k-anonymization, differential privacy, and privacy measurement [2]. These methods can to some extent protect the privacy of data, but there are also some issues. For example, k-anonymity methods are vulnerable to attribute association attacks and background knowledge attacks, while differential privacy methods may reduce the availability of data [3]. Federated Learning (FL), as an emerging machine learning framework, utilizes distributed training to enable model training without leaving the local device, effectively protecting user privacy [4]. However, classical FL frameworks are still vulnerable to privacy threats in the face of data leakage and adversarial attacks in gradient transfer operations. Although blockchain technology can well solve the node failure or malicious behavior in distributed systems, it has limitations in terms of complex node communication time, high overhead, and the inability to add or delete nodes autonomously. The research innovatively combines the two, synthesizes the advantages of both to solve these problems and reduce the risk when sharing data. This can solve the privacy protection and data security in the data sharing at the same time. Using blockchain technology as a model framework, its consensus algorithm is optimized. Then, FL is trained and updated on local data, aiming to provide a new solution in the data sharing privacy protection. The expected contribution of the study provides a theoretical foundation and practical experience for further exploring and optimizing the combination of FL and blockchain technologies in the future, which helps to promote the development and application of related technologies. The study consists of five sections in total. Section II is to analyze and summarize the research of others. Secondly, the experiment introduces how the new data sharing privacy protection model is built in Section III. Then, the performance of the model is tested in Section IV. Finally, this paper is summarized in Section V.

II. Related Works

Data sharing privacy protection is a complex and critical issue that has attracted widespread attention in the past few years. The relevant research mainly focuses on data encryption, differential privacy, and multi-party computing. Zhaofeng M et al. found that traditional centralized Internet of Things (IoT) data management solutions inevitably encountered data security challenges. In view of this, this team proposed a vehicle networking data security sharing solution that combined blockchain technology with intelligent sensors as the object. This scheme was feasible for secure sharing on LOV datasets and had advantages over traditional methods [5]. At present, there is a problem of medical data being too sensitive, making it difficult to achieve sharing in IoT data. Chen Y et al. proposed a decentralized data management method by combining blockchain technology. Under this method, users accessed and communicated data normally after verification and recording, which had a certain security and privacy [6]. To ensure the security of resource sharing in the industrial Internet, Zhang Q et al. proposed a data security sharing model for privacy protection. This model included privacy authentication, storing ciphertext indexes, and log tracking modules. This model achieved high anti-attack and data effectiveness while maintaining high-throughput data transmission, whose performance far exceeded similar models' [7]. Lv Z et al. proposed a privacy protection scheme to ensure the secure sharing of drone information to address the privacy protection of drone big data. This method had lower computational costs in key generation, encryption, and decryption, which was also superior to traditional methods [8].

FL does not need to upload data to the spatial server, thus

avoiding issues such as data privacy leakage. Nair A K et al. believed that classical FL was still vulnerable to privacy threats due to data leakage and adversarial attacks in gradient transfer operations. Therefore, this team proposed a new privacy anonymity protection framework. The central server load under this framework was reduced, while the confidentiality of shared data was increased [9]. Cho Y J et al. found that personalized FL performed well in data transmission on a single edge device. However, there were issues such as high cost and bandwidth limitations. In view of this, the team proposed a new personalized FL framework. This framework significantly reduced the heavy communication burden of large models and achieved higher testing accuracy than general models. Blockchain technology is a distributed ledger technology that ensures data security and transparency through consensus mechanisms among multiple nodes [10]. Ghotbabadi MD et al. proposed a multi-module partitioning microgrid strategy by combining blockchain technology to optimize the operation of networked microgrids for wind turbines in related environments. The operating cost of microgrids under this strategy was reduced by about 23%, and the operational reliability significantly increased [11]. Yu X et al. found that traditional multi-level security systems had the drawback of centralized authorization facilities, which made it difficult to meet the security requirements of modern distributed peer-to-peer network architectures. In view of this, this team proposed a new environmental access control model by combining blockchain technology. This model adapted well to the needs of multi-level security environments and had feasibility in practical scenarios [12].

In summary, many past studies in data sharing have also demonstrated the value of their respective applications. However, there are still some notable gaps that need to be filled in these approaches while sharing data. First, existing FL models are often limited in terms of computational and communication efficiency while protecting privacy. In addition, traditional blockchain consensus algorithms have limitations in handling node communication complexity and overhead. Second, most of the existing research focuses on the application of a single technology, e.g., using only FL or blockchain technology to address privacy protection in data sharing. However, it is often difficult to apply a single technology to simultaneously balance data privacy protection and system performance optimization. This limitation is especially obvious when dealing with large-scale and complex data sharing tasks. Therefore, an innovative approach combining FL and blockchain technologies is proposed. This combination is able to synthesize the privacy-preserving advantages of FL and the security and trustworthiness features of blockchain, while improving the efficiency and reliability of the system by optimizing the consensus algorithm. The privacy protection and security challenges in the data sharing process can be addressed more effectively through this multi-technology fusion approach.

III. Construction of Data Sharing Privacy Protection Model

Firstly, the consensus algorithm in blockchain technology is elaborated. Simultaneously, C4.5 Decision Tree (DT) is introduced for performance optimization. Finally, a novel data consensus algorithm is proposed to construct the final data sharing privacy protection model. Based on blockchain and FL, a data sharing privacy protection model targeting hospitals is constructed. Meanwhile, a reputation value calculation method is proposed to ensure privacy and security during data sharing.

A. Construction of Fault-Tolerant Mechanism Based on Improved Blockchain Consensus Algorithm

Blockchain is a distributed database technology in which data are linked together in chronological order in the form of blocks, forming an immutable chain structure [13]. Each block contains a batch of transaction records and the hash value of the previous block, ensuring the integrity and security of the entire chain in Fig. 1.

In Fig. 1, blockchain mainly consists of six modules, namely data, network, consensus, incentive, contract, and application layer. A consensus algorithm in the consensus layer is the only way to ensure that all data information in this database is tamper proof. The Practical Byzantine Fault Tolerance (PBFT) is a classic fault-tolerant consensus algorithm used to solve consensus in distributed systems in the presence of Byzantine errors, such as node failures or malicious behavior [14]. Fig. 2 shows the process of PBFT.

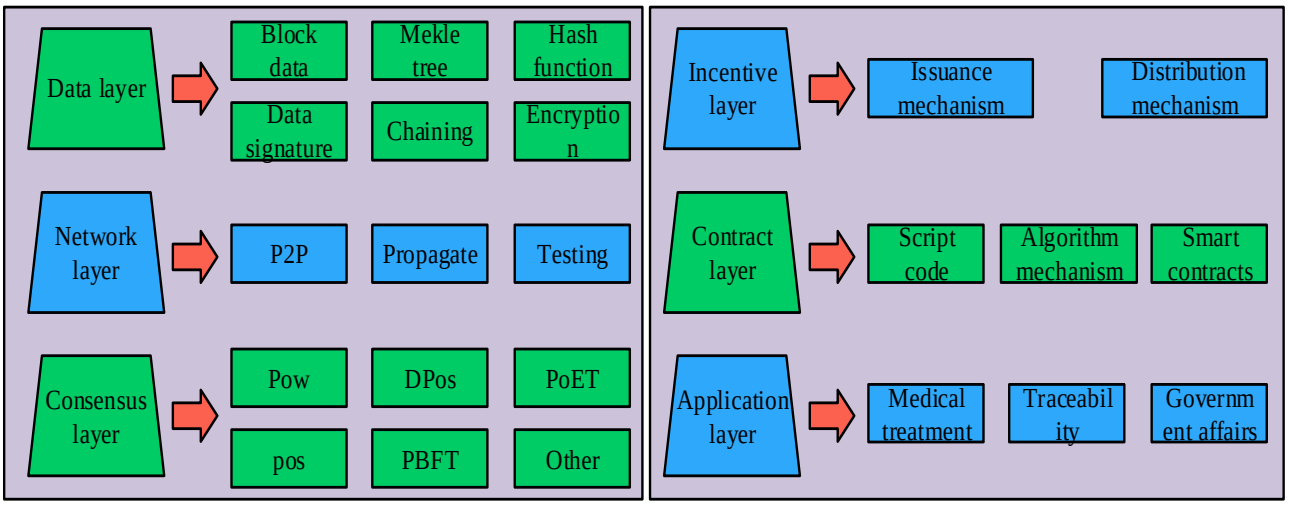


Fig. 1. Blockchain structure.

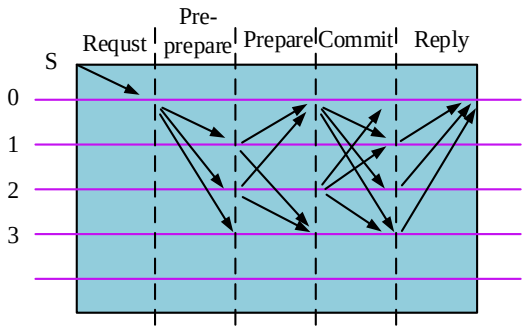


Fig. 2. PBFT algorithm process.

In Fig. 2, first, the client sends a request to the master node, which broadcasts the request to other nodes and waits for confirmation from the majority of nodes. Other secondary nodes verify and send pre prepare, prepare, and submit messages after receiving the request. When the master node receives a sufficient number of preparation messages, it submits the request and broadcasts it to other secondary nodes. Finally, all nodes accept the submission message and execute the request to ensure consensus and prevent the impact of Byzantine errors. The configuration information of each node in the blockchain when working under the same configuration information is called a view. If the master node fails, the node needs to be replaced at this time. The protocol formula for this process is represented by Eq. (1).

$$\begin{cases} V = V + 1 \\ P = V \cdot \mathrm{mod} |N| \end{cases} \tag{1}$$

In Eq. (1), V is a view number. P represents the master node number. $|N|$ refers to the nodes quantity within the blockchain system. PBFT can effectively solve the Byzantine problem through this coordination, that is, how to ensure consensus among nodes in a distributed system in the presence of faults. However, PBFT itself also has problems such as complex node communication time, high overhead, and inability to autonomously add or delete nodes. In view of this, this study introduces C4.5 DT for node optimization, which has stronger data applicability and more precise standards for handling incomplete data attributes. After each round of consensus is completed, the continuous consensus count, incorrect communication frequency, and node activity of a single node are counted in the form of reputation points. These nodes are adjusted and assigned to primary, secondary, and tertiary nodes through DT. In addition, excellent or malicious nodes are added or removed in real-time through dynamic adjustment. The view protocol of the system is changed. The first level node with high reputation is selected as a candidate node for the master node. Fig. 3 shows the entire DT-PBFT consensus algorithm.

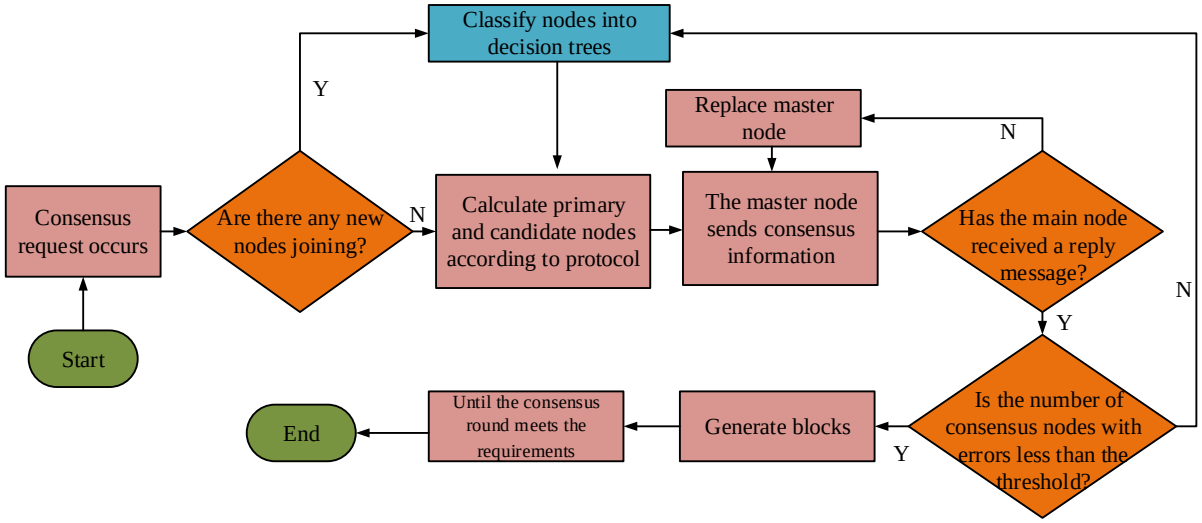


Fig. 3. Process of DT-PBFT.

In Fig. 3, first, the client makes a consensus request to the system. If a new node joins or exits at this time, it is classified through DT. If there are no new nodes, the master node and subsequent nodes are selected through the protocol, and the selected information is sent to all consensus nodes. If the master node receives preparation information that meets the threshold at this time, it determines whether the errors that occur in the consensus node are less than another threshold. If no preparation information is received, the candidate node replaces the master node and resends the information. If all requirements are met, block generation will begin until reaching the consensus count and stop. In this process, node classification is the most important step, which includes three main steps: information entropy, information gain, and gain rate calculation. Information entropy is a key indicator for measuring the categories of three types of nodes, represented by Eq. (2).

$$Info(D) = -\sum_{k=1}^{3} P_k \cdot lbP_k \tag{2}$$

In Eq. (2), P_k represents the proportion of nodes with a single category to the total nodes. $Info(D)$ represents the category information entropy of sample D. The information gain represents the degree of uncertainty of the information,

combined with the four attribute indicators in DT-PBFT, namely node reputation score, continuous consensus, downtime, and incorrect communication. The information gain at this point is represented by Eq. (3).

$$Gain(D,a) = Info(D) - \sum_{v=1}^{4} \frac{|D^v|}{|D|} Info(D^v) \quad (3)$$

In Eq. (3), a represents the feature vector. D^v represents the v th attribute indicator in sample D. The gain rate is represented by Eq. (4).

$$Gain\_ratio(D,a) = \frac{Gain(D,a)}{IV(a)} \quad (4)$$

In Eq. (4), $IV(a)$ means the characteristic vector of the fourth type of indicator error communication frequency, represented by Eq. (5).

$$IV(a) = -\sum_{v=1}^{4} \frac{|D^v|}{|D|} lb \frac{|D^v|}{|D|} \quad (5)$$

In Eq. (5), all algebraic meanings are consistent with the previous explanation. According to the above formula, these three types of nodes in C4.5 DT account for 20%, 30%, and 50% of the total, respectively. The node view switching protocol at this time is represented by Eq. (6).

$$P = V \cdot \mathrm{mod} |R_H| \quad (6)$$

In Eq. (6), $|R_H|$ represents the number of first level nodes that have completed classification sorted by reputation points. The lower the H in $|R_H|$, the lower the reputation score and the easier it is to be selected as the master node.

B. B. Construction of a Data Sharing Privacy Protection Model Combining Fault-Tolerant Consensus Mechanism and Federated Learning

This study introduces FL to continue building a shared privacy protection model after optimizing the consensus algorithm mechanism in blockchain technology. Intelligent FL gradually becomes the best choice for optimizing privacy through evolution and upgrading. Fig. 4 shows a typical FL.

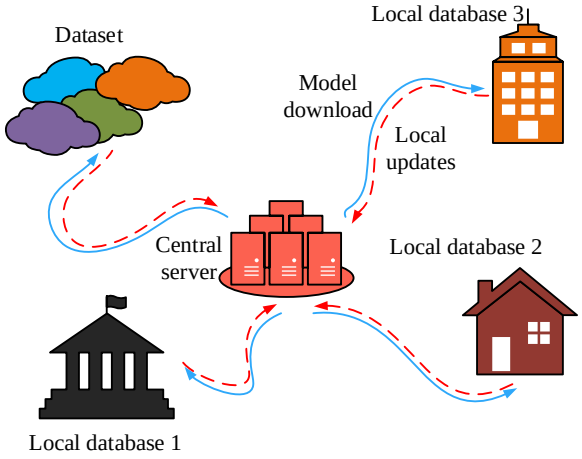


Fig. 4. Schematic diagram of FL.

In Fig. 4, the entire FL framework can be divided into three main bodies, namely the central server, participants, and local model training. Firstly, participants achieve global model training and updating by training the model locally and only sharing model parameter updates [15]. This process is then regulated and updated by the central server to reflect the local parameters of the participating parties. Meanwhile, security measures are taken to reduce communication costs and achieve model training under distributed data. However, as participants increase, the privacy leakage during model training becomes increasingly apparent. In response to this issue, this study attempts to integrate DT-PBFT with FL and proposes a novel data-sharing privacy protection model, namely DT-PBFT-FL in Fig. 5.

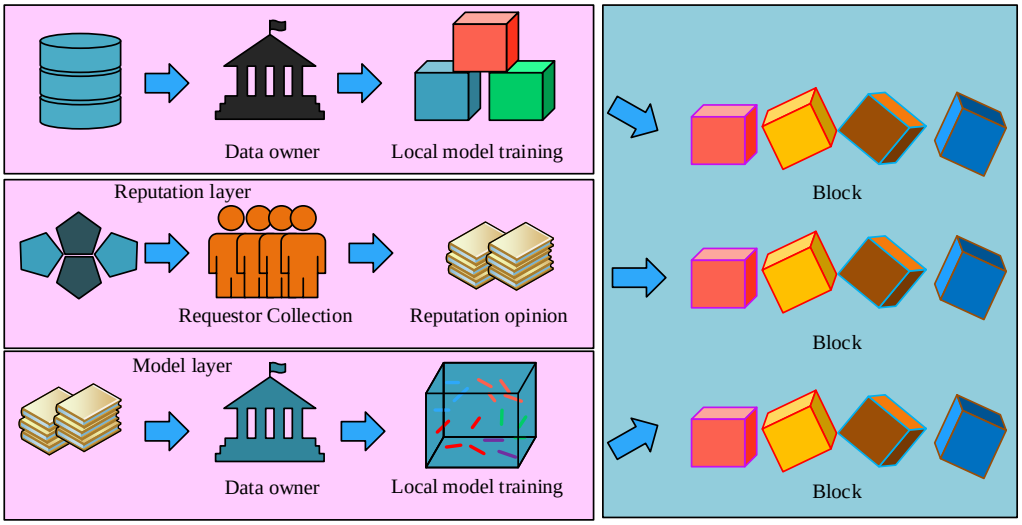


Fig. 5. Architecture of privacy protection model for medical data sharing.

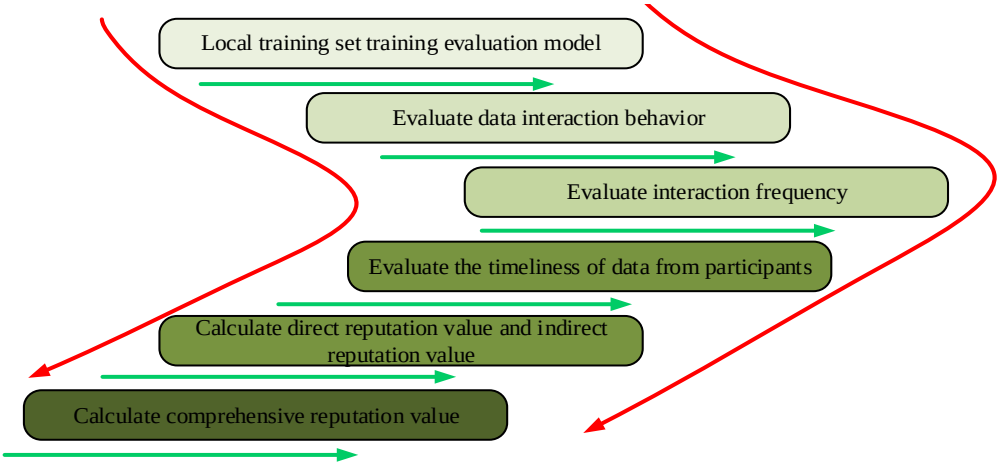


Fig. 6. Reputation calculation update.

In Fig. 5, the entire structure is divided into blockchain, reputation, and model modules. Blockchain is the foundation of the entire architecture, responsible for collecting all communication node information that responds to publishing nodes. The blockchain mainly stores the latest model training and update information. The reputation module is mainly responsible for evaluating the reputation of data publishers to ensure the quality of their published data and node stability. A model layer is mainly responsible for training FL tasks. Factors such as spatial location differences, spatiotemporal differences, and network latency can affect the task release of data owners. Therefore, this study proposes a new strategy using reputation computing to enable quantitative information exchange between data owners and requesters. Fig. 6 shows the new strategy's operational process.

In Fig. 6, the entire process is roughly divided into six stages. This includes training local models, evaluating data interaction behavior, evaluating interaction frequency, evaluating data timeliness of participants, calculating direct and indirect reputation values, calculating comprehensive reputation values and uploading them to blockchain terminals. Reputation is mainly used to evaluate the credibility of FL nodes, and its evaluation indicators have three directions: reliability, un-credibility, and uncertainty. The relationship between the three indicators is represented by Eq. (7).

$$b_{i,j} + d_{i,j} + u_{i,j} = 1 \tag{7}$$

In Eq. (7), $b_{i,j}$, $d_{i,j}$, and $u_{i,j}$ represent credibility, un-credibility, and uncertainty, respectively. Direct reputation is represented by Eq. (8).

$$\begin{cases} b_{i,j} = (1-u_{i,j})\dfrac{\alpha}{\alpha+\beta} \\ d_{i,j} = (1-u_{i,j})\dfrac{\beta}{\alpha+\beta} \\ u_{i,j} = 1-q \end{cases} \tag{8}$$

In Eq. (8), α and β represent the positive and negative times of the task publishing node and participating nodes in the event, respectively. q represents the data model's successful transmission probability. The calculation formula related to probability is represented by Eq. (9).

$$k > \eta(k+\eta = 1) \tag{9}$$

In Eq. (9), η represents a weight parameter. k represents the interaction coefficient between the model and nodes during the data sharing. The larger k, the smaller η, indicating that the reputation of the learning task publisher is affected, such as network attacks or restrictions. Indirect reputation is represented by Eq. (10).

$$\begin{cases} b_{i,j} = \sum_{e\in E} k_y b_{i,j} \\ d_{i,j} = \sum_{e\in E} k_y d_{i,j} \\ u_{i,j} = \sum_{e\in E} k_y u_{i,j} \end{cases} \tag{10}$$

In Eq. (10), e refers to the publisher of the data. E represents a collection of publishing nodes for other tasks. k_y is a weight factor of the publisher. By combining direct reputation and indirect reputation, it is convenient to store reputation through node maintenance and verify the node reputation value during data sharing, thus avoiding data loss and leakage. The comprehensive reputation is represented by Eq. (11).

$$\begin{cases} b_{i,j} = \dfrac{\sum_{y=0}^{Y} \vartheta_y \cdot b_{i,j}}{\sum_{y=0}^{Y} \vartheta_y} \\ d_{i,j} = \dfrac{\sum_{y=0}^{Y} \vartheta_y \cdot d_{i,j}}{\sum_{y=0}^{Y} \vartheta_y} \\ u_{i,j} = \dfrac{\sum_{y=0}^{Y} \vartheta_y \cdot u_{i,j}}{\sum_{y=0}^{Y} \vartheta_y} \end{cases} \tag{11}$$

In Eq. (11), ϑ_y represents the freshness decay function of a node, $\vartheta_y = Z^{Y-y}$. Z represents any number between $0-1$. $y \in (0, Y)$ represents any time period. Considering the variation of time length in data sharing, both FL task publishers and learners can affect their respective reputation values at any time.

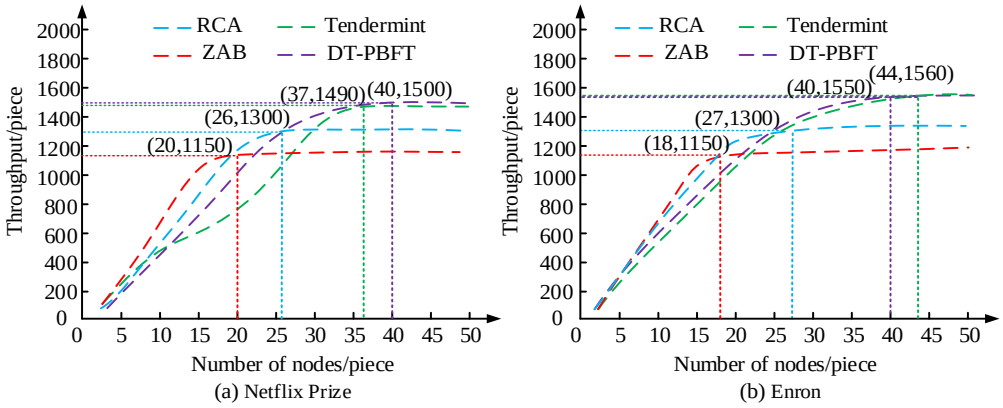
IV. Performance Testing of Data Sharing Privacy Protection Model

Firstly, multiple indicators were tested on DT-PBFT and compared with similar algorithms to verify the performance of the proposed data sharing privacy protection model. Secondly, the optimal reputation value of DT-PBFT-FL was detected. The security of different data sharing privacy protection models was compared. Finally, user evaluations were conducted.

A. Performance Testing of Block Consensus Algorithm

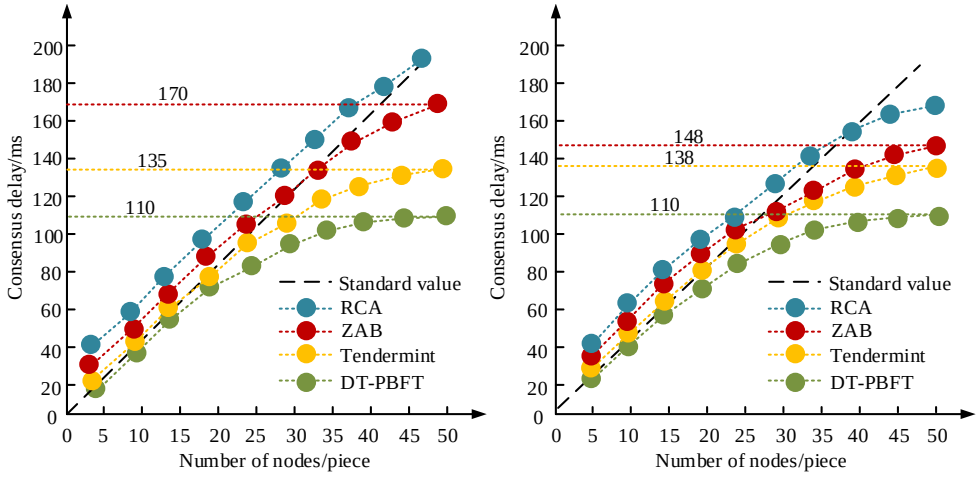
The operating system adopted Windows 10, with Intel®CoreTMi7-9700CPU@3.00GHz×32 CPU and NVIDIA GeForce RTX 1660 GPU. Hyperledger was selected as the application scenario framework. A blockchain network was built to store 100 nodes. Messages were sent through nodes, simulating Byzantine attacks. In addition, the Netflix Price and Enron datasets were introduced as data sources. Netflix Prize is a public dataset that focuses on sharing movie recommendation data information, containing nearly 20000 evaluation data from different users. Enron contains the metadata and content of millions of emails, covering communication between hundreds of users. This study compared DT-PBFT with similar popular consensus protocol algorithms: Raft Consensus Algorithm (RCA), ZooKeeper Atomic Broadcast (ZAB), and Tendermint algorithms, using throughput as a testing metric. RCA had a leader election timeout set to 150ms and a maximum batch size of 10. ZAB had a synchronization limit of 5, a time interval of 2000ms, and an initialization limit of 10. The Tendermint algorithm had a block time of 1s. Fig. 7 shows the test results.

Fig. 7(a) and 7(b) show the throughput of four algorithms on the Netflix Prize and Enron datasets. As the nodes used in blockchain networks increased, the throughput of various algorithms continued to increase and gradually stabilized in the later stages. The highest throughput of ZAB was only 1150, indicating that the transactions consensus per unit time was low and the algorithm performance was not high. The maximum throughput of DT-PBFT was 1560, which was not much different from Tendermint. However, at this point, there were only 40 nodes, which were reduced by about 4 compared to Tendermint. This study continued to test the above models based on the time difference between the initiation and completion of events in blockchain, i.e. consensus delay.



(a) Netflix Prize (b) Enron

Fig. 7. Comparison results of throughput of different consensus algorithms.



(a) Netflix Prize (b) Enron

Fig. 8. Consensus latency test results for different algorithms.

Fig. 8(a) and 8(b) show the consensus delays of four algorithms on the Netflix Prize and Enron datasets. Compared to Tendermint, DT-PBFT had a significant improvement. Especially when there were 25-30 nodes, the consensus delay was greatly reduced. The consensus delay of DT-PBFT was only up to 110ms, which was effectively reduced by 28ms compared to Tendermint's 138ms. The above data indicated that the proposed algorithm was more suitable for current data sharing work and had excellent computational performance. Finally, this study used consensus rounds as a variable and set the initial number of Byzantine nodes to 300. The security of the above four algorithms and similar algorithms in references 7 and 8 were tested using the Byzantine nodes in the system as an indicator. The test results are shown in Table I.

In Table I, after 8 iterations, the lowest number of Byzantine nodes in RCA was 196. The minimum number of Byzantine nodes for ZAB after 8 iterations was 172. The minimum number of Byzantine nodes in Tendermin after 8 iterations was 139. The minimum number of Byzantine nodes in reference 7 was 151 after 8 iterations in similar approaches, while the minimum number of Byzantine nodes in reference 8 was 142. The proposed DT-PBFT had a minimum of 130 Byzantine nodes after 8 iterations. In summary, the proposed method effectively removed the Byzantine nodes in the system, reducing the probability of Byzantine nodes being selected as the main node and ensuring the security of the entire system. The Netflix Prize dataset and the Enron dataset had significant differences in the performance of each model under the comparison test due to the different uniformity of data distribution, different feature complexity, and different data noise outliers. For data types that were distributed, with high security requirements, multiple nodes, and frequent data updates, the algorithm was able to give full play to its advantages and provide an efficient and secure data sharing solution.

B. *Performance Testing of Data Sharing Privacy Protection Model*

This study used the software environment of Python 4.0 to test the proposed DT-PBFT-FL data sharing model on the platform of Python 3.9. The Cerner Health Facts dataset was used as the testing data source. This dataset contains clinical data from multiple hospitals, including over 40000 pieces of information on patient diagnosis, treatment, medication prescriptions, and more. The training set and test were divided in an 8:2 ratio, with 50 initial nodes and weight parameters $k = 0.4$ and $\eta = 0.6$. This study first determined the reputation threshold within the 0-1 range to determine the optimal state of DT-PBFT-FL in Fig. 9.

TABLE I. COMPARISON OF BYZANTINE NODE REMOVAL RESULTS USING DIFFERENT ALGORITHMS

Data set	Algorithm	Consensus round							
		1	2	3	4	5	6	7	8
Netflix Prize	RCA	285	274	263	250	242	239	218	204
	ZAB	282	273	264	251	237	219	204	185
	Tendermin	280	261	242	227	201	186	164	139
	Reference 7	283	276	263	241	219	186	174	152
	Reference 8	288	271	253	237	221	198	179	165
	DT-PBFT	279	264	240	221	200	173	152	131
Enron	RCA	286	271	261	248	229	210	206	196
	ZAB	285	270	254	232	211	203	189	172
	Tendermin	280	267	246	231	214	197	172	158
	Reference 7	281	266	247	227	204	187	163	151
	Reference 8	279	264	232	209	187	164	153	142
	DT-PBFT	279	254	236	204	183	164	143	130

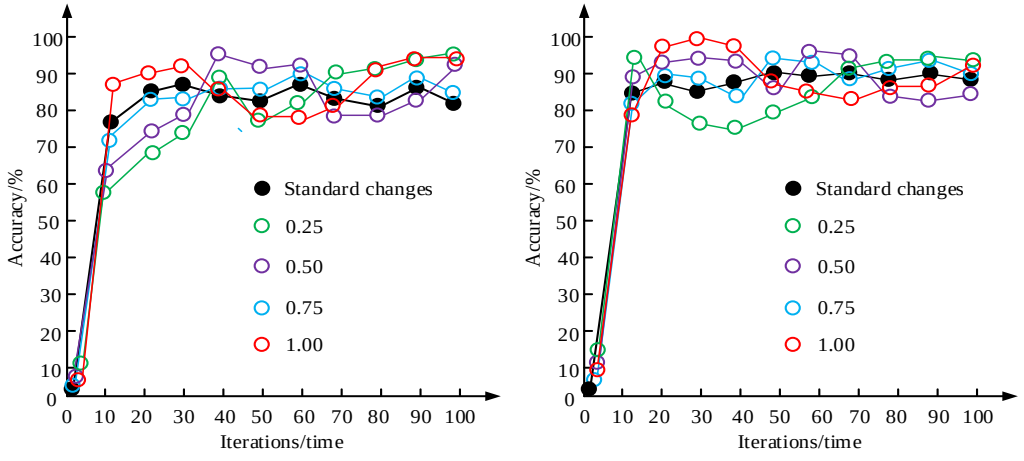


Fig. 9. Changes in model accuracy under different reputation values.

Fig. 9(a) shows the model training accuracy results for four different reputation values on the training set. Fig. 9(b) shows the model training accuracy results for four different reputation values on the test set. After introducing reputation values in the early stages of iterative changes, the training accuracy of the entire model was significantly improved. The higher the reputation value, the better the training effect. However, not a reputation value of 1.0 performed the best. When the reputation value was 0.75, the model training accuracy changed more in line with the standard change line, indicating that the training data were completely reliable and the data quality was better. At this time, the highest accuracy of DT-PBFT-FL was 96%. On this basis, this study introduced Role-Based Access Control (RBAC), Peer Trust Model (PTM), Eigen Trust Model (ETM), Reputation-based Trust Management (RTM), Fuzzy Reputation Model (FRM), and Web of Trust (WoT) of the same type. The initial roles for the RBAC algorithm were set to 4. The maximum number of maximal privileges was set to 10. The trust threshold for PTM was 0.6. The maximum number of interactions was 50. The trust decay rate for ETM was set to 0.1. The reputation decay rate for the RTM was 0.05. The depth of trust propagation for WoT was 3. The maximum number of nodes was 20. Similar methods in the literature were also introduced, i.e., methods in references 9 and 10. In the experiment, reputation value was used as the testing indicator and 10 iterations were conducted. The model reputation after each iteration was recorded in Table II.

In Table II, in the first 5 iterations, the reputation changes of the 7 data sharing models were relatively small. However, in the subsequent 5 iterations, various models demonstrated significant node updating capabilities. The magnitude of change in the values of the more popular methods of the same type was greater and more pronounced than those detected in studies 7 and 8, especially after the 8th iteration. Relatively speaking, DT-PBFT-FL had the largest change in reputation value, with the lowest reputation value of 0.32. This indicated that the greater the magnitude of data changes, the greater the model ability to detect malicious nodes. If malicious nodes disguised themselves as normal data nodes in the first 5 iterations, none of the 7 models exhibited excellent diagnostic capabilities. Therefore, within a certain iteration range, the proposed model had superiority and feasibility. The study validated the medical clinical information data in the Cerner Health Facts dataset using the effectiveness of malicious node detection as an indicator. The results are shown in Fig. 10.

Fig. 10 (a) shows the outlier detection results of RBAC, Fig. 10(b) shows the outlier detection results of WoT, Fig. 10(c) shows the outlier detection results of the model proposed in study 10, and Fig. 10(d) shows the outlier detection results of the proposed model. From Fig. 10, the outlier detection results of RBAC and the model proposed in study 10 were poor as the test samples increased. Although the outlier detection of WoT had greater similarity with the standard value, there were still some data samples with lower outlier detection. The detection efficiency and effectiveness of the proposed method matched the standard values to a high degree, which indicated that the data transmission security of DT-PBFT was improved to a greater extent by combining FL. Finally, to explore the actual differences between the proposed new model and the WoT with the best data performance, the study used stability, safety, economy, and data validity as test indicators. Customer evaluations were scored through random selection. The maximum score was 100 points, and the passing score was 60 points. Fig. 11 shows the scoring results.

Fig. 11(a) and Fig. 11(b) show the customers' rating results for WoT and DT-PBFT-FL. The rating range for WoT from 4 clients was between 70-90, while the evaluation scores for the proposed model were concentrated at 90 or above. The highest stability score was 97, safety was 95, data validity was 97, and economy was 94. Comparing the average scores of the two models, the average score of WoT was 80.5, and the average score of DT-PBFT-FL was 93.8%. In summary, from customer evaluations, the proposed model had better overall performance and was more popular than similar models.

TABLE II. THE REPUTATION VALUE AFTER 10 ITERATIONS

Iterations/time	1	2	3	4	5	6	7	8	9	10
RBAC	0.75	0.75	0.74	0.73	0.73	0.68	0.64	0.58	0.54	0.50
PTM	0.75	0.75	0.74	0.73	0.73	0.69	0.64	0.61	0.55	0.51
ETM	0.74	0.74	0.74	0.73	0.71	0.68	0.64	0.57	0.52	0.48
RTM	0.75	0.75	0.73	0.73	0.72	0.67	0.65	0.62	0.57	0.52
FRM	0.75	0.75	0.75	0.73	0.72	0.68	0.66	0.61	0.57	0.52
WoT	0.75	0.75	0.73	0.73	0.72	0.67	0.63	0.52	0.44	0.41
Reference 9	0.75	0.75	0.74	0.72	0.71	0.68	0.65	0.64	0.59	0.53
Reference 10	0.75	0.74	0.73	0.72	0.69	0.68	0.64	0.61	0.55	0.48
DT-PBFT-FL	0.74	0.73	0.73	0.72	0.7	0.52	0.45	0.4	0.32	0.34

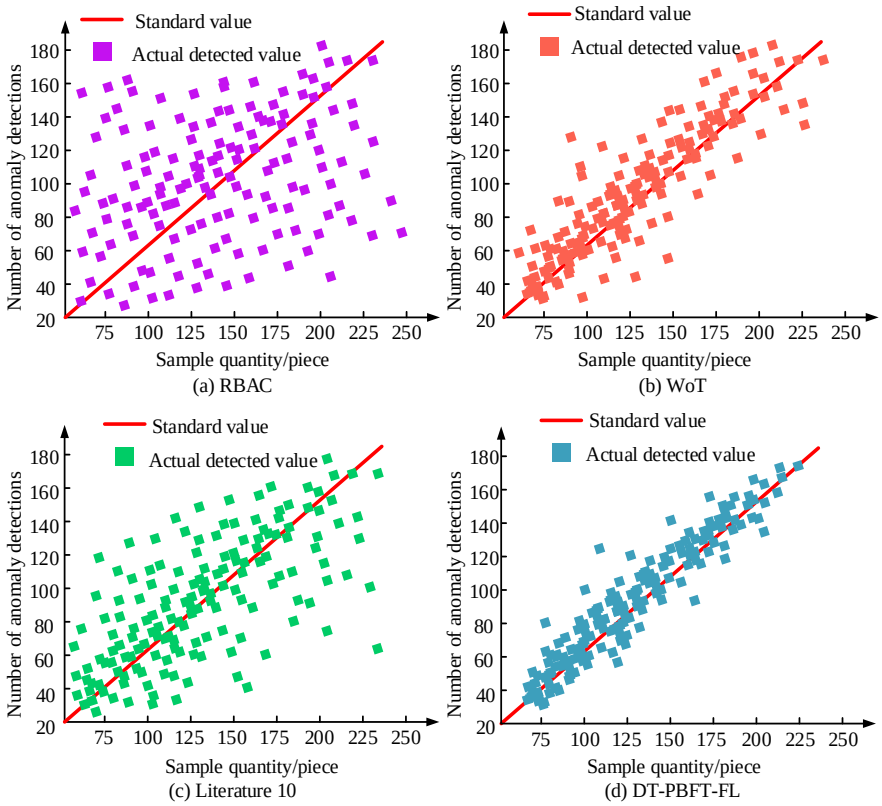


Fig. 10. Test results of anomalous data detection for four model.

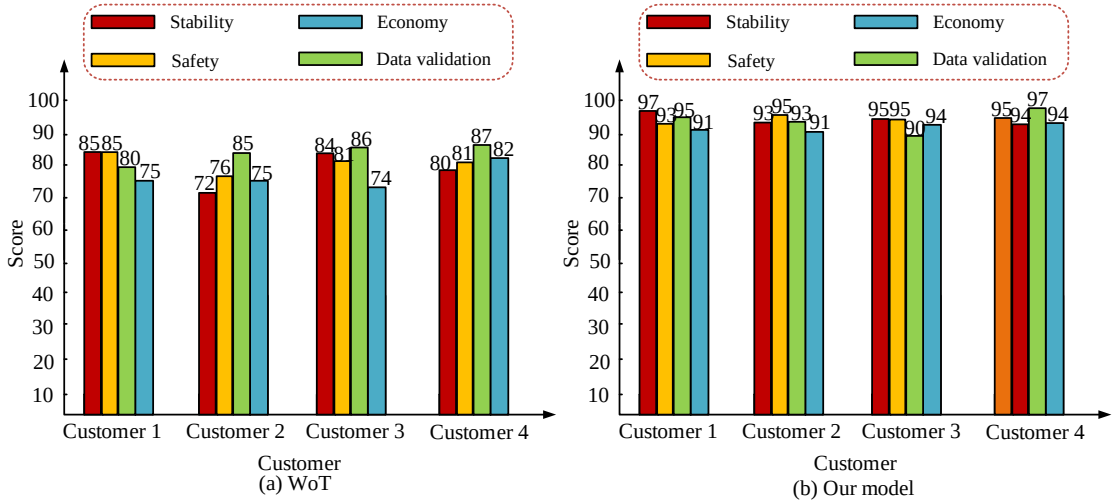


Fig. 11. Customer rating results for two models.

V. Conclusion

Data sharing has significant significance in the era of big data, but privacy protection is a major challenge in data sharing. Therefore, first, this study took blockchain technology as the framework, analyzed the PBFT consensus algorithm, and introduced C4.5 DT for optimization. After completion, local data model training was achieved through FL. An optimized PBFT was used for data sharing services and data supervision. Finally, a new data sharing privacy protection model, DT-PBFT-FL, was proposed. The maximum throughput of DT-PBFT was 1560, and the number of nodes at this time was only 40. The consensus delay of DT-PBFT was up to 110ms, which was effectively reduced by 28ms compared to Tendermint's 138ms. After 8 iterations with an initial 300 Byzantine nodes, the minimum number of Byzantine nodes in DT-PBFT was 130. In addition, when the reputation was 0.75, the training accuracy of DT-PBFT-FL was more in line with the standard change curve, and the model performance was optimal. After 10 consecutive iterations, the reputation value of this model was as low as 0.32. After user evaluation, the stability score was the highest at 97, the safety was the highest at 95, the data validity was the highest at 97, and the economy was the highest at 94.

In summary, the proposed DT-PBFT-FL data sharing privacy protection model can complete current data sharing tasks with high standards, and has high feasibility and stability. Future research can further explore the model's performance overhead and scalability to enhance its effectiveness in more practical application scenarios.

ACKNOWLEDGMENT

The research is supported by National Key R&D Program "Confidential computing hardware acceleration technoloty" (No.2023YFB4503200).

REFERENCES

[1] Jia B, Zhang X, Liu J, Zhang Y, Huang K, Liang Y. Blockchain-enabled Federated Learning Data Protection Aggregation Scheme with Differential Privacy and Homomorphic Encryption in IIoT. IEEE Transactions on Industrial Informatics, 2021, 18(6):4049-4058.

[2] Ahmed F, Wei L, Niu Y, Zhao T, Zhang W, Zhang D, Dong W. Toward fine-grained access control and privacy protection for video sharing in media convergence environment. International journal of intelligent systems, 2022, 37(5):3025-3049.

[3] Xu Z, Luo M, Kumar N. Privacy-Protection Scheme Based on Sanitizable Signature for Smart Mobile Medical Scenarios. Wireless Communications and Mobile Computing, 2020, 2020(1):1-10.

[4] Li Y, Lu Y, Qi S, Zheng Y, Chen X. Cpds: Enabling Compressed and Private Data Sharing for Industrial Internet of Things Over Blockchain. IEEE transactions on industrial informatics, 2021, 17(4):2376-2387.

[5] Zhaofeng M, Lingyun W, Weizhe Z. Blockchain-Driven Trusted Data Sharing with Privacy-Protection in IoT Sensor Network. IEEE Sensors Journal, 2020, 21(22):25472-25479.

[6] Chen Y, Meng L, Zhou H, Xue G. A Blockchain-Based Medical Data Sharing Mechanism with Attribute-Based Access Control and Privacy Protection. Wireless Communications and Mobile Computing, 2021, 2021(5):1-12.

[7] Zhang Q, Li Y, Wang R, Liu L, Tan Y, Hu J. Data security sharing model based on privacy protection for blockchain-enabled industrial Internet of Things. International Journal of Intelligent Systems, 2020, 36(1):94-111.

[8] Lv Z, Qiao L, Hossain M S, Choi B J. Analysis of Using Blockchain to Protect the Privacy of Drone Big Data. IEEE Network, 2021, 35(1):44-49.

[9] Nair A K, Sahoo J, Raj E D. Privacy preserving Federated Learning framework for IoMT based big data analysis using edge computing. Computer Standards and Interfaces. 2023, 86(8):1-20.

[10] Cho Y J, Wang J, Chirvolu T, Joshi G. Communication-Efficient and Model-Heterogeneous Personalized Federated Learning via Clustered Knowledge Transfer. IEEE journal of selected topics in signal processing, 2023, 17(1):234-247.

[11] Ghotbabadi M D, Dehnavi S D, Fotoohabadi H, Mehrjerdi H, Chabok H. Optimal operation and management of multi-microgrids using blockchain technology. IET Renewable Power Generation, 2022, 16(16):3449-3462.

[12] Yu X, Shu Z, Li Q, Huang J. BC-BLPM: A Multi-Level Security Access Control Model Based on Blockchain Technology. China Communications, 2021, 18(2):110-135.

[13] Kalapaaking A P, Khalil I, Rahman M S, Atiquzzaman M, Xun Y, Almashor M. Blockchain-Based Federated Learning With Secure Aggregation in Trusted Execution Environment for Internet-of-Things. IEEE transactions on industrial informatics, 2023, 19(2):1703-1714.

[14] Guo S, Zhang K, Gong B, Chen L, Ren Y, Qi F, Qiu X. Sandbox Computing: A Data Privacy Trusted Sharing Paradigm Via Blockchain and Federated Learning. IEEE Transactions on Computers, 2023, 72(3):800-810.

[15] Gheisari M, Hamidpour H, Liu Y, Saedi P, Raza A, Jalili A, Rokhsati H, Amin R. Data Mining Techniques for Web Mining: A Survey. Artificial Intelligence and Applications, 2023, 1(1): 3-10.