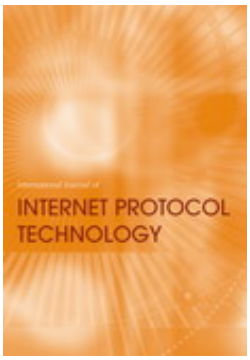


International Journal of Internet Protocol Technology

ISSN online: 1743-8217 - ISSN print: 1743-8209
https://www.inderscience.com/ijipt

Enhancing data security in massive data sets using blockchain and federated learning: a loosely coupled approach

Haiyan Kang, Bing Wu

DOI: 10.1504/IJIPT.2024.10068232

Article History:

Received:	31 July 2024
Last revised:	02 October 2024
Accepted:	02 October 2024
Published online:	06 January 2025

Enhancing data security in massive data sets using blockchain and federated learning: a loosely coupled approach

Haiyan Kang* and Bing Wu*

Department of Information Security,
Beijing Information Science and Technology University,
Beijing, China
Email: kanghaiyan@126.com
Email: wubing1150274764@163.com
*Corresponding authors

Abstract: The properties of the blockchain are not suitable for the storage of sensitive privacy data, and the excellent characteristics of the blockchain will be seriously affected by the existence of massive amounts of data on the chain. To address the above issues, propose a Loosely Coupled Local Differential Privacy Blockchain Federated Learning method (LL-BCFL). First of all, a client selection mechanism is put forward to ensure honesty and positivity in joining the training client and the correct effectiveness of the final global model aggregation. Secondly, use federated learning to alleviate the 'data silos' phenomenon and achieve joint training of big data stored in distributed multi-parties. Additionally, a differential privacy method is introduced to act on federated learning networks to avoid inference attacks. Finally, the MNIST dataset was used to confirm the availability of the LL-BCFL method on balanced and unbalanced datasets.

Keywords: federated learning; blockchain; differential privacy; massive data processing.

Reference to this paper should be made as follows: Kang, H. and Wu, B. (2024) 'Enhancing data security in massive data sets using blockchain and federated learning: a loosely coupled approach', *Int. J. Internet Protocol Technology*, Vol. 17, No. 1, pp.31–41.

Biographical notes: Haiyan Kang is a Senior Member of the China Computer Federation (No. E200028533M), ACM Membership (No. 9495204) and Member of Privacy Protection Committee of China Confidentiality Association. He received his PhD degree in Computer Application Technology from Beijing Institute of Technology, China in 2005. His research interests include information system security, privacy preserving and Natural Language Processing (NLP). He is currently working as a Professor in the Department of Information Security at Beijing Information Science and Technology University, Beijing, China.

Bing Wu is currently studying for a Master's degree in Electronic Information at Beijing Information Science and Technology University. Her research interests include privacy protection in the field of machine learning and data compliance utilisation.

This paper is an expanded version of paper 'Research on federated sharing methods for massive data in blockchain' presented at '7th EAI International Conference on Smart Grid and Internet of Things', TaiChung, Taiwan PRC, 18–19 November 2023.

1 Introduction

In the context of the big data era, data is both the foundation and the key. It can promote major breakthroughs in the development of advanced technologies and industries in big data industries such as the internet, finance, artificial intelligence and cloud computing, China's data volume has been increasing year by year, and it is the world's largest data producer. Every day, it obtains massive amounts of data from e-commerce platforms, transportation systems, hospitals and other major fields, China ranks first in the world in terms of data contribution, and the rate of data growth has only increased year by year. Data from authoritative organisations show that China's new data volume was 7.6 ZB in 2018, and it is expected that China's new data volume will reach 48.6 ZB in 2025, accounting for 27.8% of the total global data volume, with a compound annual growth rate of up to 30%, making it the world's largest data circle.

The existence of massive data provides a basic guarantee for the development of advanced technology in China, but data security problems such as data leakage and data theft have become increasingly serious. Didi Company excessively

collects user information and the exposure of over 540 million records of Facebook subscriber on Amazon cloud services, and the promulgation of data security-related legal documents, i.e., the 'Cybersecurity Law of the PRC' and the 'Data Security Law of the PRC', data security has gradually been paid close attention by the state, enterprises and individuals, which to a certain extent makes it impossible for massive data to play its best role. In order to solve the problem of safe use of massive amounts of data, this paper uses a federated learning mechanism (Tang et al., 2024) that allows multiple participants to collaborate on training models without sharing their own data. Each participant has direct management and control over its own data, ensuring the privacy and security of the data. At the same time, the cooperation of multiple participants makes the data set more diverse, greatly improving the generalisation ability and accuracy of the model.

By leveraging the security features of blockchain, such as decentralisation and independent autonomy, which can effectively achieve secure storage of distributed data (Wang et al., 2023). However, storing big data on the blockchain will lead to higher costs and slower efficiency. In response to the problem of 'data silos' caused by the limitation of blockchain storage and the decentralised nature of massive data storage, use 'chain up and chain down' ideas is proposed, combined with federated learning technology to accomplish safe and high efficient processing of the huge volume of data. With the deepening of research, many scholars have proposed a new idea of introducing federated learning mechanisms in the application of blockchain technology to jointly build models. For example, Kang et al. (2019) ensured the accuracy of reputation calculation with the help of blockchain, and realise more secure and reliable model training with the help of federated learning combined with incentive mechanism.

Federated Learning (FL) can effectively address privacy and security issues in machine learning, but the risk of privacy leakage cannot be ignored. Federated learning ensures the security of training by combining perturbation mechanism and encryption method. Among them, the perturbation mechanism is achieved with the help of Differential Privacy (DP). Noise is used to directly perturb the initial data to protect privacy data. Zhang et al. (2023a) proposed a grid clustering method AGCluster based on LDP. By applying local differential privacy technology to clustering problems, the clustering quality is improved while ensuring data safety. The encryption method indirectly achieves privacy protection during data exchange by combining cryptographic tools and other security technologies. Common security technologies include homomorphic encryption and secret sharing. Zhang et al. (2020) proposed a key-value data collection method LDPKV based on encryption algorithm and LDP to solve the problem of effective processing of key-value data without sharing private information. Yu and Chen (2023) proposed a federated aggregation scheme to address the communication overhead problem in model training, and introduced homomorphic encryption technology to ensure data security. Zhang et al. (2023b) proposed an adaptive federated deep learning algorithm for non-independent homogeneous data in combination with homomorphic encryption, which ensures the privacy of the training data while improving the training speed of the model.

The above related research has made significant progress in data security and processing, but there are still three problems that need to be solved urgently, namely, (1) the low efficiency and difficulty in managing distributed data terminal information retrieval in federated learning under massive data, (2) there is a possibility that dishonest and inactive clients will join the federated learning network, (3) the validation of the experiment is only for balanced data sets and lacks universality. After conducting careful research and deep analysis on the above three problems, the key innovations of this article are summarised as below:

1 Propose a blockchain oriented credible federal study LL-BCFL method, Loosely Coupled Local Differential Privacy Blockchain Federated Learning, which accomplish the safe processing and effective utilisation of big data.

2 Propose a Client Selection Mechanism (CSM) that selects reliable clients to join the training and ensure high accuracy of the final model.

3 Design a LDP mechanism for parameter transfer in federated learning, which addresses privacy leakage issues during model training.

4 Test the validity of the LL-BCFL method on the MNIST data set.

2 Related study

2.1 Federated learning

Federated learning (Dwork and Lei, 2009) combines multiple data sources to train models without uploading local data. Combining with technologies in multiple fields such as deep learning and privacy protection, which aims to solve the issue of utilises efficiently big data. Here is the definition of FL.

Federated learning (Dwork and Lei, 2009). Define N joiners $\{F_1,...,F_N\}$ has a corresponding data set of $\{D_1,...,D_N\}$ and jointly train the global model M_{FED}. Let the accuracy of the machine learning model be V_{SUM} and the accuracy of the FL model be V_{FED}, then the accuracy loss is:

$$\left|V_{FED}-V_{SUM}\right|<\delta \tag{1}$$

Among them, the non-negative real number δ is a non-negative real number. In ideal conditions, the value of δ is very small.

Federated learning has effectively addressed 'data silos' and data safety issues, but it also has shortcomings in reliability. Consequently, combining relevant privacy protection technologies to solve the safety issues of

parameter transfer, the veracity and reliability issues of joining model training clients, are important challenges faced by federated learning.

2.2 Blockchain

Blockchain is a decentralised ledger technology with safety properties, for instance, transparency, immutable and undeniable, which has been widely applied in various fields, for instance, medical care and transportation. Blockchain technology has the above significant advantages, but its security and scalability problems cannot be ignored.

Blockchain ensures safety through cryptography and consensus algorithms, but the security mechanism has theoretical weaknesses, making it vulnerable to malicious information attacks, resource abuse attacks and so on. Low throughput, high data load, and low query efficiency are the causes of scalability problems in blockchain systems.

Using blockchain for the huge volume of data storage may lead to issues such as low efficiency in transactions, queries due to data overload. Therefore, consider combining it with federated learning to avoid storing massive amounts of data on the blockchain and at the same time solve the problem of effective and safe utilisation of distributed storage data.

2.3 Local differential privacy technique

Differential privacy relies on randomisation algorithms in 2008. It can effectively prevent differential and inference attacks. This paper adopts relaxed DP, and below is its definition.

(ε, δ)-*Differential privacy* (Wang et al., 2023). Given n users, for any random algorithm M, take any two adjacent data sets D and D' as inputs, let any subset output by algorithm M be Y $(Y \in R)$, then

$$Pr\left[M(D) = Y\right] \leq e^{\varepsilon} \times Pr\left[M(D') = Y\right] + \delta \tag{2}$$

Then, the algorithm M satisfies (ε, δ)-DP. Among them, δ is the relaxation parameter. Parameter ε represents the degree of privacy protection. If higher privacy protection is required, a smaller ε value can be set.

Differential privacy often uses perturbation mechanisms, for instance, Gaussian mechanism and Laplace mechanism to add random noise to the original data for perturbation. Among them, the Gaussian mechanism accomplish LDP by adding Gaussian noise with a mean of 0 and a variance of $\sigma^2 I$ to the output result $f(t)$, that is, $M(t) = f(t) + M(\sigma^2 I)$. The Laplace mechanism adds Laplace noise to the output result $f(u)$ to accomplish LDP, that is, $M(t) = f(u) + Laplace(\Delta f / \varepsilon)$. Among them, $Laplace(\bullet)$ is Laplace noise.

3 Design and analysis of method LL-BCFL

3.1 Description of the problem

Large amounts of data on the blockchain may affect the performance of the blockchain's own characteristics. The distributed storage huge data will aggravate the 'data silos' issue, making it impossible to effectively use massive data. The distributed nature of federated learning can resolve the datastore and usage issues of centralised learning. However, also have some problems such as malicious clients launching privacy attacks that affect model training operations (Wu and Kang, 2024).

Therefore, the realisation of compliant utilisation and secure sharing of massive data requires the above problems need to be solved. Table 1 provides a description of notations and parameters.

Table 1 Notations and parameters

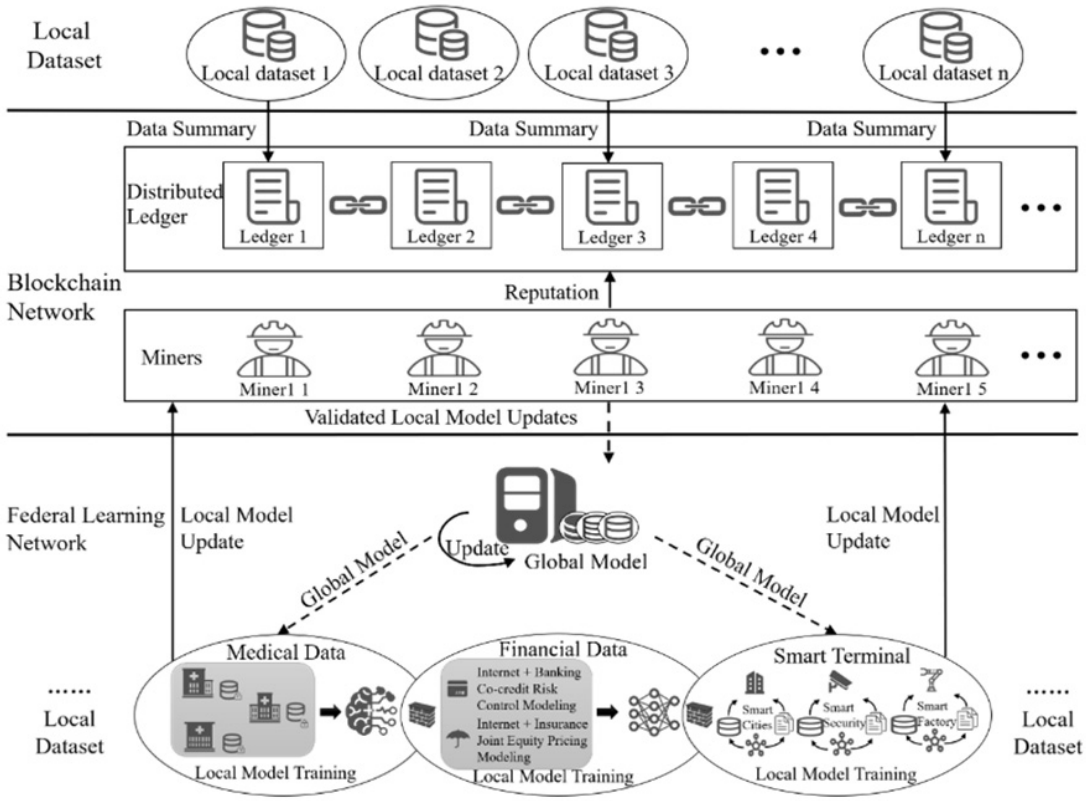
Notation	*Meaning*
M N	Quantity of clients Quantity of FL entrants
T	Total quantity of FL rounds
ε	Privacy budget in LDP definition
δ	Relaxation parameter in LDP definition
w_i	Model parameters
u_i	Local model trained on the client side

3.2 Architecture of the method

In response to the above issues, this article proposes a LL-BCFL method to realise the safe and effective use of big data. The LL-BCFL method adheres to the idea of 'chain up and chain down' for construction, its architecture is shown in Figure 1.

The data digest is stored on the blockchain, while massive data is stored and processed off the chain. The method mainly includes the following three sub methods: (1) loosely coupled BCFL method for integrating blockchain technology and federated learning mechanisms, (2) FL method based on LDP to effectively deal with possible inference attacks in federated learning and guaranteeing data security in the process of model training parameter transmission and (3) big datastore and processing method for solving the issue of effective use of large amounts of data.

Figure 1 System architecture diagram



The core points of the LL-BCFL massive data efficient processing method designed are described as below. Firstly, the method ensures the security of the summary information storage on the chain and the reliability of the data during the model training process by means of the transparency, traceability and tampering characteristics of the blockchain itself, as well as the verification update and reputation calculation operations of the miners in the BCFL. Secondly, using FL technology to accomplish the 'integration' of the huge volume of data from multiple sources stored in various terminals, effectively alleviating the phenomenon of 'data silos' while improving the accuracy and applicability of the learning model. Besides, a LDP mechanism is introduced to impede possible inference attacks during training to make sure the implementation of trusted federated learning. The method proposed in this article can be applied to industries with extremely high data privacy requirements such as medical care, finance and security and provides a valid solution to the use of big data.

3.2.1 Loosely coupled BCFL method

Using the loosely coupled FL method (Wang et al., 2023) to integrate blockchain and FL. The construction method is shown in Figure 1. Among them, federated learning adopts a server client construction approach, mainly used for training models and joining the network by uploading model parameters to the blockchain. By using blockchain verification models to update the correctness of parameters, calculate and store the reputation value of the client, and join the network by issuing reputation values to FL.

Using the distributed ledger in the blockchain network, message include brief data summaries and reputation values of the clients are recorded to ensure the traceability and tamperability of the relevant data. The distributed ledger sends the reputation value to the central server during each round of training to provide a basis for it to choose the clients to join training, which improves the efficiency and performance of model training while standardising the behaviour of the clients. With the help of miners in the blockchain network, data support is provided for the client selection mechanism of federated learning, that is, miners play the role of a selector and select appropriate clients for model training. The process of selecting a client can be described as below:

1 *Client self-evaluation*: The client assesses its own data based on the required data specifications and computing resources, which mainly includes the evaluation of data size, data type, etc.

2 *Server assessment*: The server uses the Subjective Logic Model (SLM) method to obtain the client's reputation index, and combine the self-evaluation value and historical reputation value of the client to obtain the comprehensive reputation value of the client.

3 *Select the client*: based on comprehensive reputation value, the client is selected as the participant in the federation learning model training.

The implementation process of the CSM is shown in Algorithm 1.

Algorithm 1: Select_Client

Input: Quantity of clients M, quantity of participants in federated learning model training N, total quantity of FL rounds T.

Output: List $client\_select\_list$ of clients that join the federated learning model training.

1. Define the list $client\_eval\_list$, $client\_score$, $client\_select$.
2. **for** $i \leftarrow 1$ to M **do**
3. Iterate through M clients, weight and sum each client's assessment of its own data size, quality and category to get the client self-assessment value $client\_eval$, and add it to the list of $client\_eval\_list$.
4. $client\_eval \leftarrow s_1 * w_1 + s_2 * w_2 + s_3 * w_2$
5. **if** $t \leftarrow 1$ **then**
6. Select N clients to participate in the model training based on their evaluation values from largest to smallest and add them to the $client\_select$ list.
7. **else**
8. Obtain the reputation value $client\_rep\_list$ from the list $client\_rep$ of client's historical reputation value, add it to the weighted self-evaluation value to get the client's comprehensive assessment value, and add it to the $client\_rep\_list$ list.
9. $client\_rep \leftarrow client\_eval * w_4 + client\_rep * w_5$
10. Select N clients to join training based on their comprehensive evaluation values and add them to the $client\_select\_list$ list.
11. **end for**
12. **return** $client\_rep\_list$

The clients that join the model training in each round must be determined through a client selection mechanism. This cycle is repeated round by round until the model converges. The server selects reliable participants with high reputations who hold high-quality data sets for federated learning tasks to bring significant improvements in learning efficiency. Utilising miners to provide validation mechanisms for updating local models in federated learning. The following is a description of the validation update process:

1 *Get updated parameters*: After the client uploads the local model update, the miner receives and processes it.

2 *Verify model parameters*: After receiving local model updates, miners validate them based on rules and constraints established.

The realisation method of the verification mechanism is described in Algorithm 2.

Algorithm 2: Valid_Model

Input: Updated client local model u_i, test data set $valid\_data$, number of correctly predicted labels cor_num.

Output: Client local model accuracy acc.

1. Prediction on updated local model u_i using validation data set.
2. $prediction \leftarrow u_i.predict(valid\_data)$
3. Iterate through the prediction results and real labels, record the number of correctly predicted labels using accuracy as a performance metric, and thus judge the effectiveness of local model updates.
4. **if** $prediction == valid\_data$:
5. $cor\_num += 1$
6. $acc = cor\_num / len(valid\_data)$
7. **return** acc

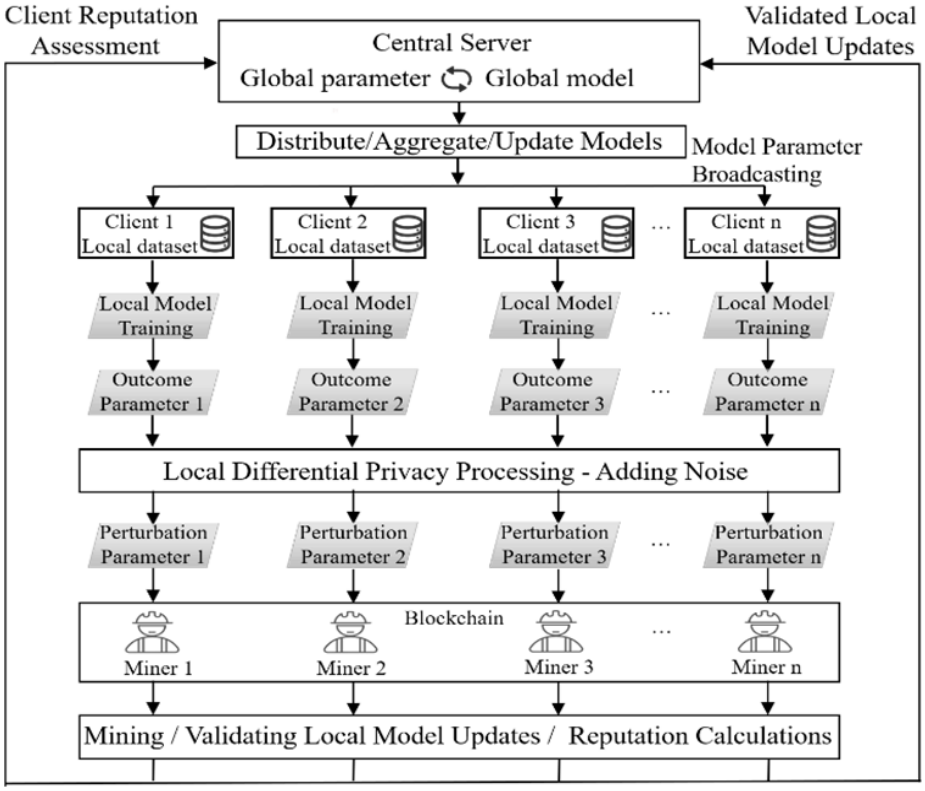
Miners will send the updated and verified model parameters to the server. This method of effectively controlling the model parameters participating in the global aggregation process can avoid the impact of malicious or invalid parameters on the availability of the BCFL model.

3.2.2 A FL method based on LDP

This article utilises a LDP mechanism to address the issue of inference attacks that may occur during the intermediate parameter exchange process in federated learning. Figure 2 shows the architecture of this method. The complete training process of this method can be summarised below.

The central server mainly completes the initialisation, distribution, aggregation of global parameters and updating of the global model. In cross-device federated learning, there are problems such as communication bandwidth limitations, differences in data quality of data sets and differences in the training quality of clients, which require the selection of clients to join training. In the first round of training, the server determines the participating clients based on each client's own evaluation value, using a selection method from largest to smallest. In each subsequent round of training, the server chooses the clients based on the reputation evaluation value of each client by miners in the blockchain. The client selection mechanism improves the quality and efficiency of model training, while encouraging clients to honestly and actively join the learning network. After the selected client obtains the global parameters, it locally trains a federated learning model based on LDP. Each client participating in the training will obtain its corresponding result parameters after the training is completed, add noise for disturbance, and send it to the miners for transaction verification. At the same time, miners also need to complete the calculation of reputation values and the evaluation of client reputation and upload them to the server. Finally, the server completes the global model aggregation. The above process is repeated step-by-step until the model converges.

Figure 2 Federated learning architecture based on LDP



3.2.3 Massive data storage and processing method

Massive datastore mainly includes two parts: namely (1) simple summary data and (2) each terminal data set. Among them, the simple summary data on the blockchain are stored in the distributed ledger, which mainly includes the descriptive information of the local user, the summary information such as the relevant information of the data set data, and the reputation value that the miner evaluates for each participant after each round of training. The data or information recorded in the distributed ledger cannot be tampered with, which can effectively improve the security of the data. At the same time, the public data recorded in the distributed ledger is authorised for users to retrieve it, which can increase the training efficiency. The data set data required for model training is always stored locally at each terminal and is directly stored and managed by the data holder.

This system completes the data processing process mainly includes three parts: (1) client local model training and introduction of LDP noise, (2) the miner verifies the updated local model parameters after training and (3) the server aggregates local model updates that pass validation.

Some data is stored on the blockchain, which effectively utilises the transparency and immutability of the blockchain, facilitates reputation calculation, client retrieval and other operations and increases learning efficiency. The off-chain storage of massive data ensures that each end-user has the authority to manage the data, and under the premise of ensuring data security, it fully and effectively utilises the data of all parties to train the model in a distributed manner, which effectively alleviates the phenomenon of 'data silos' and improves the accuracy and universality of the model.

3.3 Analysis of method LL-BCFL

The LL-BCFL method utilises blockchain to store relevant data summary information, and the security of on chain information can be ensured by the unique properties of the blockchain. Secondly, possible inference attacks on process parameters during model training can be effectively avoided through local differential privacy mechanisms. In addition, high quality clients joining model training are ensured by the client selection mechanism. consequently, the proposed LL-BCFL method has higher security.

The number of participants is M and the overall number of iterations be T. The time complexity of the aggregation algorithm is $O\left(\log\left(M\right)\right)$, and the time complexity of the LL-BCFL method is $O\left(T\log\left(M\right)\right)$ and its value is equal to T times $O\left(\log\left(M\right)\right)$.

4 Experiment and analysis

4.1 Experimental data set

The experiment selected the MNIST data set include 60,000 training samples and 10,000 test samples for training and testing. The MNIST data set includes 10 types of 28×28 grey-scale images of handwritten digits. Experimental simulation of horizontal federated learning process, loading MNIST data set through code automation, providing training data for model training participants.

The data used in federated learning comes from different terminals and has differences due to the characteristics of the domain industry and other factors, and this further affects the accuracy and effectiveness of the final model. Consequently, the system balances the data samples before model training begins. For the balanced data set (Tang et al., 2023), a set amount of Independent Identically Distribution (IID) random data is randomly selected for each user terminal as its model training data set before the training starts to obtain better training results. The processing for the unbalanced data set (Tang et al., 2023) is realised by expanding the number of users. That is, in order to reduce the influence of unbalanced data on the model training results, the actual number of users is first expanded to twice the original number, and then random selection and balancing are performed.

4.2 Experimental environment

Hardware use Intel(R) Core(TM) i5-8265U CPU @1.60 GHz. Memory is 8 GB. The Windows 10 (64-bit) operating system was used to train the model using Python 3.8 in the pycharm environment.

Use pytorch 1.7.1 to train a deep learning model and add differential privacy noise. This experiment uses the convolutional neural network structure to perform iterative training of the global model. Use Stochastic Gradient Descent (SGD) to adjust model parameters and optimise local models.

4.3 Model validity assessment experiment

This part evaluates the validity of Method LL-BCFL, considering two situations: balanced data sets and unbalanced data sets. Mainly focusing on the variables of noise type, the quantity of local users and training rounds, this study explores their impact on the final accuracy of the model. The implementation methods of the three experiments on the MNIST data set are below.

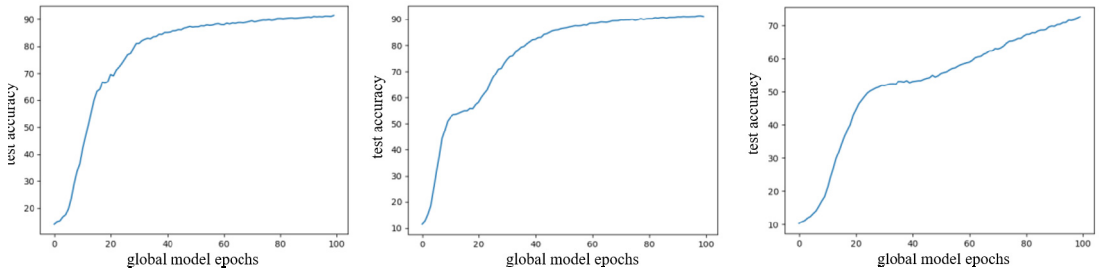
Both experiments were set up as follows: the differential privacy parameter C = 10%, the quantity of training rounds is 100, and the added noise in each group of experiments was Laplace noise, Gaussian noise and the no-DP was used as a control.

(1) Noise type serves as a parametric variable in the experiment. (Set the quantity of FL entrants to $N = 100$)

1) Figure 3 shows the results of training the model using the balanced data set.

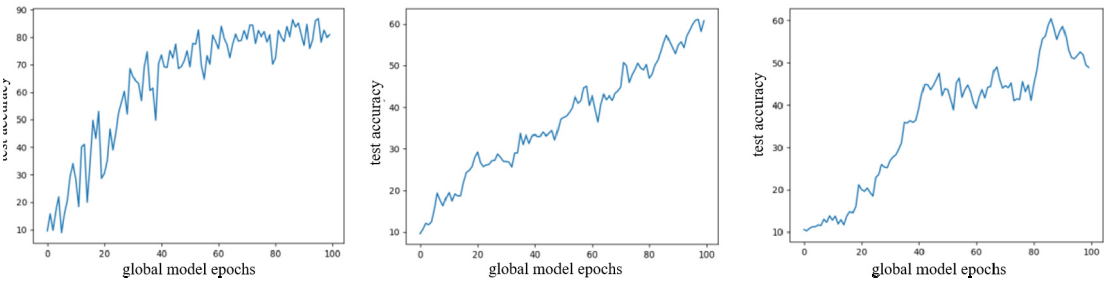
2) Figure 4 shows the results of training the model using the unbalanced data set.

Figure 3 Global accuracy under IID (rounds is set to 100) (see online version for colours)



(a) No noise introduced (b) Laplace noise is introduced (c) Gaussian noise is introduced

Figure 4 Global accuracy under non-IID (rounds is set to 100) (see online version for colours)



(a) No noise introduced (b) Laplace noise is introduced (c) Gaussian noise is introduced

From the Figures 3 and 4, the following conclusions can be obtained:

1) Under the premise that the quantity of participants is the same, a balanced data set is used for training for 100 rounds, adding noise will have a negative effect on the accuracy of model training. Among them, after adding Gaussian noise, the model accuracy reached a maximum of only 70%, which had a great impact on the model accuracy. Nevertheless, adding Laplace noise the final model accuracy reached 90%. That is, when training with a balanced data set for fewer rounds, the effect of adding Laplace noise on the model accuracy is much lower than the effect of adding Gaussian noise on the model accuracy.

2) Under the premise that the quantity of participants is the same, the unbalanced data set is used for training for 100 rounds, compared to Figure 3, the curve fluctuates greatly and the final accuracy has decreased. After adding noise, there is also a situation where convergence cannot be achieved. Additionally, the addition of two kinds of noise all appear to be unable to converge. Therefore, it is impossible to accomplish the training of the model by adding noise in the unbalanced data set under the case of fewer training rounds.

(2) The quantity of local users serves as a parametric variable in the experiment.

The quantity local users are set to 100, 150 and 200 to obtain the influence of the quantity of users participating in training on the model accuracy.

1) Under the above conditions, set the quantity of local users to 100. The results of training with a balanced data set are shown in Figure 3. The results of training with an imbalanced data set are shown in Figure 4.

2) Under the above conditions, set the number of local users to 150.

Based on Figures 5 and 6, combined with Figures 3 and 4, can be summarised as below:

(1) As shown in Figure 5, the accuracy of the model with the addition of Laplace noise grows slowly, and the final accuracy only reaches 67% and does not converge. Compared with Figures 3(b) and 5(b) shows that the higher the quantity of participants, the more training rounds are needed to complete the learning. The accuracy of federated learning with Gaussian noise added in Figure 5(c) finally converges at 90%, compared to the results of adding Gaussian noise in Figure 3(c), it can be seen that the more local users join the training, the higher the accuracy of the final model.

(2) As can be seen from Figure 6, the final accuracy is lower than that of Figure 5. Compared to the results of Figure 3 trained with balanced data set and 100 local users, the accuracy of the no-DP and Laplace DP methods decreases, while the accuracy of the Gaussian DP method increases.

3) The number of local users is set to 200.

(1) By comparing Figures 7 and 8 with Figures 5 and 6, it can be seen that increasing the number of users while keeping the number of training rounds unchanged will cause a decrease in model accuracy.

(2) The convergence curves of the FL accuracies under both the Laplace and Gaussian mechanisms under the comparison of different quantity of local users involved in model training using balanced and unbalanced data sets illustrate that the accuracy of the model decreases with the increase in the quantity of local users involved in federated learning training, and that a highly accurate model cannot be obtained. Therefore, a balance should be found in the parameter settings of the number of training rounds and the number of local users to ensure a higher accuracy model.

(3) The quantity of training rounds serves as a parametric variable in the experiment.

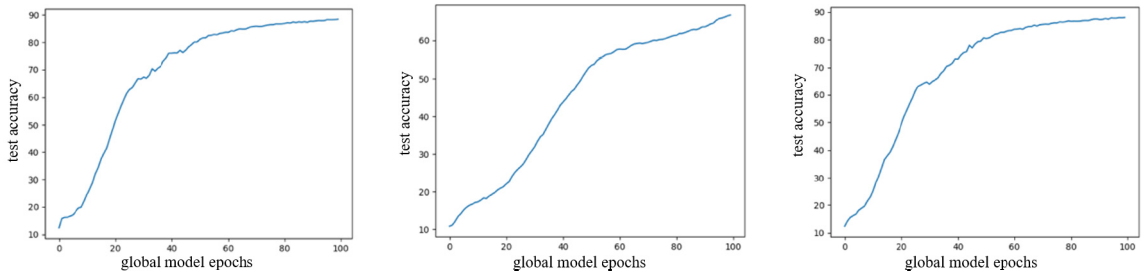
Set the training rounds to 100, 200 and 500, respectively.

1) Setting the quantity of training rounds to 100. The results of training with a balanced data set are shown in Figure 3. The results of training with an unbalanced data set are shown in Figure 4.

2) Setting the quantity of training rounds to 200.

Based on Figures 9 and 10, combined with Figures 3 and 4, can be summarised as below:

Figure 5 Global accuracy under IID (local user is set to 150) (see online version for colours)



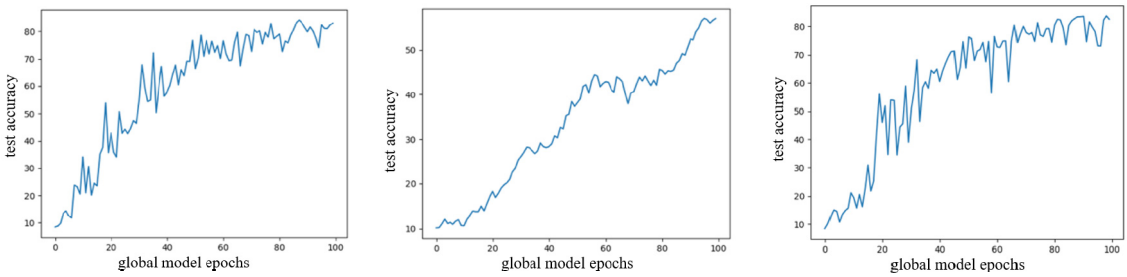
(a) No noise introduced (b) Laplace noise is introduced (c) Gaussian noise is introduced

(1) Available from Figure 9, the accuracy curves of the global model after adding two types of noise shows a similar trend of change, and the accuracy of the trained models is between 80% and 90%.

(2) As shown in Figure 10, the final accuracy of the model has decreased compared to Figure 9. Compared to the FL curve in Figure 4(b), the final accuracy of Figure 10(b) has increased by 10%. The federated learning curve in Figure 10(c) shows an obvious rise in training results compared to Figure 4(c). That is, the accuracy of the model trained with an unbalanced data set will continue to improve as the number of training rounds increases.

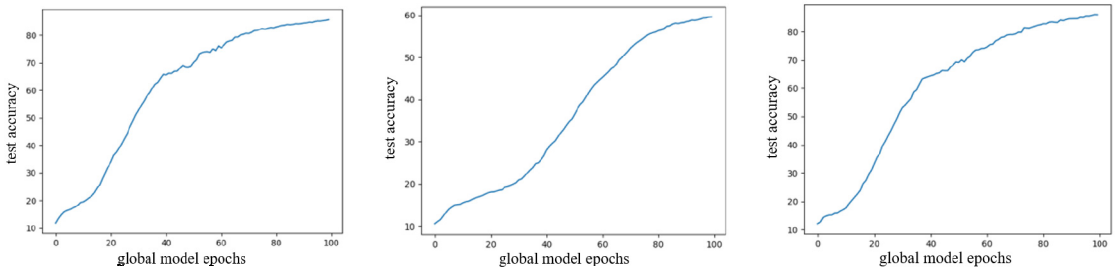
3) Setting the quantity of training rounds to 500.

Figure 6 Global accuracy under non-IID (local user is set to 150) (see online version for colours)



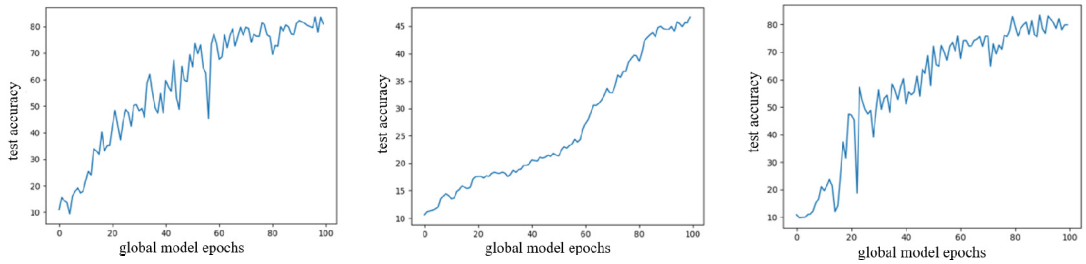
(a) No noise introduced (b) Laplace noise is introduced (c) Gaussian noise is introduced

Figure 7 Global accuracy under IID (local user is set to 200) (see online version for colours)



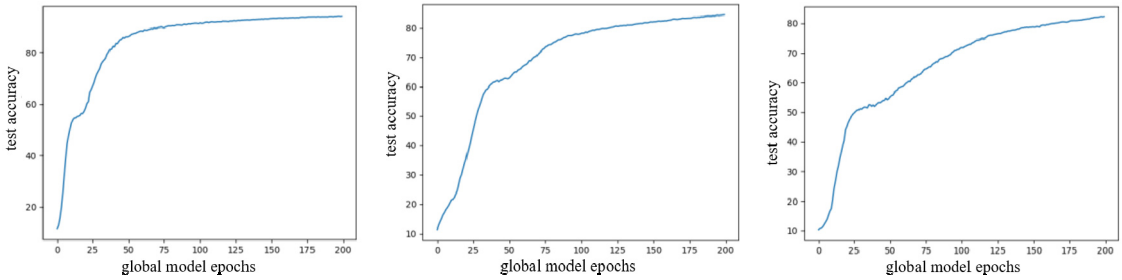
(a) No noise introduced (b) Laplace noise is introduced (c) Gaussian noise is introduced

Figure 8 Global accuracy under non-IID (local user is set to 200) (see online version for colours)



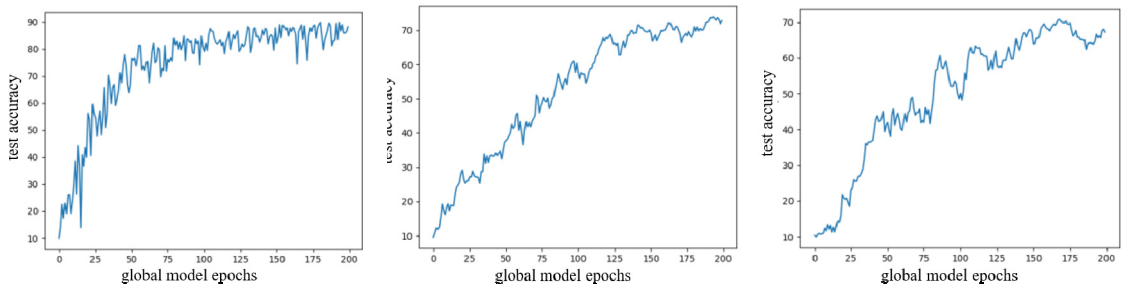
(a) No noise introduced (b) Laplace noise is introduced (c) Gaussian noise is introduced

Figure 9 Global accuracy of IID (rounds is set to 200) (see online version for colours)



(a) No noise introduced (b) Laplace noise is introduced (c) Gaussian noise is introduced

Figure 10 Global accuracy of non-IID (rounds is set to 200) (see online version for colours)



(a) No noise introduced (b) Laplace noise is introduced (c) Gaussian noise is introduced

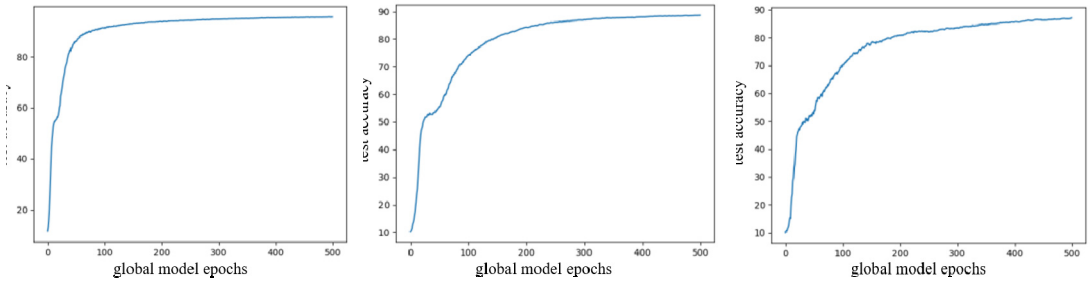
Based on Figures 11 and 12, combined with Figures 3, 4, 9 and 10, can be summarised as below:

(1) Available from Figure 11(b) and 11(c) that the final model accuracy is very close, and both of them have the characteristics of using no-difference privacy technology for model training in the early stages of training.

(2) From Figure 12, compared to Figure 10, the final accuracy of the model has improved, and compared to Figure 9, the final accuracy of the model is basically the same.

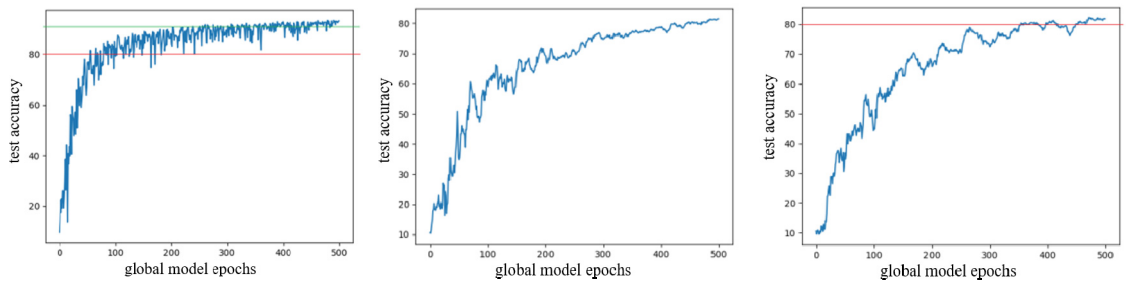
(3) Using unbalanced data sets and setting different numbers of training rounds, through comparison of accuracy convergence curves, it can be seen that as long as there are enough rounds of federated learning training, the commonly existing unbalanced data sets in reality can also be well aggregated to obtain high-precision models. Use a balanced data set and set different training rounds. By comparing the accuracy convergence curves, the accuracy of the model sustainable growth as the quantity of training rounds increases. But it is worth noting that how to use unbalanced data sets to obtain a higher accuracy model in fewer training rounds is the focus of the next research.

Figure 11 Global accuracy of IID (rounds is set to 500) (see online version for colours)



(a) No noise introduced (b) Laplace noise is introduced (c) Gaussian noise is introduced

Figure 12 Global accuracy of non-IID (rounds is set to 500) (see online version for colours)



(a) No noise introduced (b) Laplace noise is introduced (c) Gaussian noise is introduced

5 Conclusion

This article proposes and designs a LL-BCFL method, to achieve compliant utilisation of big data. Blockchain storage summary information, to improve the retrieval efficiency of user related information. Use FL to solve the 'sharing' problem, and use the LDP to effectively avoid possible inference attacks during training. Design a client selection mechanism to select clients that have been verified by miners and have higher reputation values to enhance the efficiency of model training and the accuracy of the aggregation model. Finally, for the balanced data set and the unbalanced data set, the original undifferentiated private FL was used as a comparison to verify the availability of the LL-BCFL method on the MNIST data set. Future work will be devoted to studying the combination of federated learning and advanced technology industries, and privacy protection technology, so as to obtain a global model with strong security and high accuracy.

Acknowledgement

This project is partially supported by the Beijing Advanced Innovation Centre for Future Blockchain and Privacy Computing Fund (No. GJJ-23), and National Social Science Foundation, China (No. 21BTQ079).

References

Haiyan, K. and Jie, D. (2023) 'A cross encryption scheme for data security storage in cloud computing environment', *International Journal of Internet Protocol Technology*, Vol. 16, No. 1, pp.1–1.

Kang, H., Yuanrui, J. and Shuxuan, Z. (2009) 'Enhanced privacy preserving for social networks relational data based on personalized differential privacy', *Chinese Journal of Electronics*, Vol. 31, No. 4, pp.741–751.

Tang, L.T., Wang, D. and Liu, S.Y. (2023) 'Data augmentation scheme for federated learning with non-IID data', *Journal on Communications*, Vol. 44, No. 1, pp.164–176.

Tang, X., Liang, Y. and Chen, W. (2024) 'Multi-stage federated learning mechanism with non-IID data in internet of vehicles', *Journal of Computer Research and Development*, Vol. 61, No. 9, pp.2170–2184.

Wang, L.P., Guan, Z. and Li, Q.S. et al. (2023) 'Survey on blockchain-based security services', *Journal of Software*, Vol. 34, No. 1, pp.1–32.

Wu, B. and Kang, H. (2024) 'Research on federated sharing methods for massive data in blockchain', *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering Smart Grid and Internet of Things*, pp.12–27.

Yu, S.X. and Chen, Z. (2023) 'Efficient secure federated learning aggregation framework based on homomorphic encryption', *Journal on Communications*, Vol. 44, No. 1, pp.14–28.

Zhang, D.Y., Ni, W.W., Zhang, S., Fu, N. and Hou, L.H. (2023) 'A local differential privacy based Priva-cy-preserving grid clustering method', *Chinese Journal of Computers*, Vol. 46, No. 2, pp.422–435.

Zhang, X.J., Fu, N. and Meng, X.F. (2020) 'Key-value data accurate collection under local differential privacy', *Chinese Journal of Computers*, Vol. 43, No. 8, pp.1479–1492.

Zhang, Z.F., Li, Q.D. and Fu, Y. et al. (2023) 'Adaptive federated deep learning with non-IID data', *Chinese Journal of Computers*, pp.1–13.