Distributed Machine Learning Privacy Protection Algorithm for Tax Big Data Analysis

Ying Fang, Ling Pu*, Ning Zhang, Jun Liang and Shaojuan Ouyang

Department of Economics, Qinhuangdao Vocational and Technical College, Qinhuangdao 066100, China
E-mail: plpuling@163.com
**Corresponding Author*

Received 22 September 2025; Accepted 05 November 2025

Abstract

With the acceleration of informatization and digitization, the tax system has generated massive amounts of data with diverse types and large scales. However, the data sharing across regions and institutions faces challenges on privacy protection and compliance. Therefore, a distributed machine learning privacy protection algorithm for tax big data is proposed, and a multi-layer secure transmission mechanism combining differential privacy, homomorphic encryption, and secure multi-party computation is designed. In the experiment, real invoices and tax declaration data from provincial tax bureaus, as well as simulated data generated based on these data, are selected to compare various existing methods. The results showed that the accuracy in classification tasks reached 0.87, which was 2.35%–8.75% higher than that of traditional distributed methods. In the regression task, the mean square error and mean absolute error were reduced by 10%–40% and 22.03%, respectively. Compared to homomorphic encryption methods, the designed

Journal of Cyber Security and Mobility, Vol. 15_1, 67–94.
doi: 10.13052/jcsm2245-1439.1513

method reduced communication overhead by 59.67% and achieved a fault tolerance of 96.38% under the 10% node dropout rate. In addition, the accuracy decrease in poisoning attack scenarios was only 25.29%, which was superior to other methods. This algorithm can achieve high predictive performance and robustness while ensuring privacy and compliance, providing effective technical support for intelligent tax governance.

Keywords: Tax big data, distributed machine learning, privacy protection, homomorphic encryption, secure multi-party computation, robustness, compliance.

1 Background

With the acceleration of informatization and digitization, the tax system has accumulated a massive amount of data with a large scale and diverse types. These data cover invoice circulation, tax declaration, transaction records, and cross-departmental information sharing, collectively referred to as tax big data [1]. Tax big data not only reflect the macro situation of economic operation, but also contain the potential value of identifying risks, evaluating credit, and optimizing tax governance [2]. In recent years, Distributed Machine Learning (DML) has gradually become an important technological path for intelligent analysis of tax big data due to its advantages in multi-node parallel computing, heterogeneous data fusion, and complex model training. Extensive research has been conducted on the cleaning, storage, and analysis methods of tax data, with related work mainly focusing on invoice anomaly detection, taxpayer credit evaluation, and anti-tax avoidance modeling, laying the data processing foundation for the subsequent DML [3, 4]. Wahab and Bakar integrated descriptive and predictive analysis to address the low compliance rates and large amounts of data that were difficult to diagnose in Malaysia's digital economy income tax. In the prediction phase, a single classifier was introduced along with three integrated methods of packaging, boosting, and voting, as well as grid evolution parameter optimization, thereby accurately identifying potential non-compliant taxpayers and optimizing tax administration [5]. Ullah et al. proposed a Quantile-on-Quantile regression to address the unclear role and policy gaps of market-based economic tools in ecological sustainability. By analyzing the asymmetric relationship between environmental taxes and ecological sustainability in the seven major green economies from 1995 to 2018, environmental taxes significantly improved ecological quality [6]. In response to the increasingly

complex global tax evasion methods and low recognition rates of traditional rule systems, Ariyibi et al. proposed an Artificial Intelligence (AI) anti-fraud framework that integrated Machine Learning (ML), Deep Learning (DL), and natural language processing to optimize the fairness and transparency of tax regulation [7].

However, the sensitivity and compliance requirements of tax data pose serious privacy and security challenges for its computation and sharing in cross-institutional and cross-regional environments [8, 9]. Scholars both domestically and internationally have conducted extensive research on DML and privacy protection. Y. Tan et al. proposed an intelligent tax system based on AI and ML to address the traditional tax auditing relying on manual labor, low efficiency, and easy errors. The system achieved real-time data analysis, risk identification, and automated compliance processes [10]. S. Liu et al. proposed a privacy-preserving distributed depth deterministic policy gradient algorithm to solve the low hit rate of edge cache due to the difficulty in ensuring user privacy in mobile edge computing networks. Cache optimization was transformed into a distributed model free Markov decision problem and predicted popularity through federated learning, thereby improving cache hit rates to better than baseline methods and protecting user privacy throughout the entire process [11]. Padmanaban proposed a five-dimensional privacy protection framework that covered authorization management, access control, data security, network integrity, and scalability to address the privacy leakage in the context of AI and blockchain integration. By integrating encryption, de-identification, multi-layer ledger, and k-anonymity technology, the privacy protection efficiency and security were optimized [12]. Zhu et al. proposed an AI penetration supervision framework to address the low efficiency and difficulty in identifying hidden risks in administrative prosecutorial supervision under massive financial data. ML, natural language processing, and network analysis were integrated to achieve real-time monitoring and adaptive models, thereby optimizing case handling and cross-border law enforcement collaboration [13].

In summary, although existing research has made positive progress in tax data analysis, distributed learning architecture design, and privacy protection mechanisms, there are still significant shortcomings. Firstly, most methods are focused on finance or healthcare fields, lacking systematic adaptation to the complexity and compliance requirements of tax big data. Secondly, some work focuses on privacy protection and neglects the data integrity and robustness in the distributed training process. In addition, existing solutions often face high computational and communication costs

in large-scale, multi-source heterogeneous tax data environments. Therefore, how to improve the efficiency and reliability of DML while ensuring data privacy and compliance remains a core challenge that urgently needs to be addressed. In response to this research gap, a DML with Privacy Protection (DML-PP) algorithm framework for tax big data is proposed, aiming to achieve a balance between security, performance, and scalability, and provide technical support for intelligent tax governance. The research innovatively combines the gradient and parameter transmission mechanism of privacy protection, providing new ideas for the privacy and security of tax big data.

This research focuses on the privacy and efficiency challenges faced in the cross-regional and cross-institutional sharing and modeling of tax big data. A DML-PP algorithm that integrates differential privacy, homomorphic encryption, and Secure Multi-party Computation (SMPC) is proposed. This approach overcomes the bottleneck of traditional models in balancing data security and computational performance, enabling efficient collaborative training without leaking the original data. The algorithm improves accuracy by 2.35% to 8.75% in classification and regression tasks, reduces communication overhead by 59.67%, and achieves a node fault tolerance rate of 96.38%, demonstrating excellent security, stability, and scalability. This research not only provides technical support for smart tax governance, but also offers a viable path for building a compliant and efficient collaborative tax data analysis system.

The overall structure of the study consists of four sections. The first section proposes a distributed learning method framework for tax scenarios and designs a privacy-preserving gradient and parameter transmission mechanism. In the second section, an experimental environment and dataset are constructed to systematically evaluate the proposed algorithm on accuracy, efficiency, and communication overhead. The third section discusses the advantages and limitations of the method from security, robustness, and policy compliance. The fourth section summarizes the research results and looks forward to further development directions.

2 Methodology

Firstly, a DML framework for tax big data is constructed to achieve collaborative modeling in a multi-source heterogeneous data environment through local training, parameter uploading, and global aggregation. Secondly, to address potential inversion and theft risks during gradient and parameter transmission, a privacy-preserving transmission mechanism is proposed by

incorporating differential privacy, homomorphic encryption, and SMPC to form a multi-layer defense chain. Differential privacy mitigates single-sample leakage risks, homomorphic encryption ensures secure aggregation in the ciphertext domain, and SMPC eliminates single-point trust vulnerabilities. These three mechanisms work synergistically to effectively defend against Gradient Inversion Attacks (GIA) and Model Reconstruction Attacks (MRA), ultimately forming the DML-PP algorithm.

2.1 Distributed Machine Learning Framework for Tax Big Data

With the comprehensive promotion of the "Golden Tax Phase IV" project, China's tax system is accelerating towards a data-driven intelligent governance stage. In the process of integrating multidimensional data such as invoice flow, fund flow, logistics and contract flow, tax authorities gradually construct a digital portrait covering the entire lifecycle of taxpayers. This transformation has not only improved the efficiency of tax collection and management, but also triggered a tense relationship between taxpayers' privacy rights and public interests. Especially in scenarios such as cross-departmental data sharing, tax enterprise data interaction, and third-party platform access, the sensitivity, recognizability, and traceability of tax data have significantly increased, putting traditional "de-identification" methods at risk of failure. From the perspective of governance logic, the application of tax big data is not purely a technical process, but a social and technological practice embedded in specific institutional structures, power relations, and compliance frameworks [14]. As data controllers, tax authorities need to balance "regulatory effectiveness" and "protection of taxpayer rights" when exercising their collection and management powers. In addition, taxpayers' information rights, control rights, and remedies in data use have not been fully protected at the legal and institutional levels, resulting in a structural disadvantage when facing algorithmic tax collection and management. Therefore, privacy protection is not only a technical issue, but also an important component of governance legitimacy. When faced with such complex tax big data, the traditional single node centralized computing method can no longer meet the dual requirements of modeling efficiency and security compliance [15]. Therefore, building an efficient, scalable, and privacy-preserving DML framework has become an important technical support for intelligent tax governance. Unlike traditional centralized modeling, this framework not only needs to solve the storage and computation for large-scale tax data, but also needs to ensure that data does not leave the domain and privacy is not

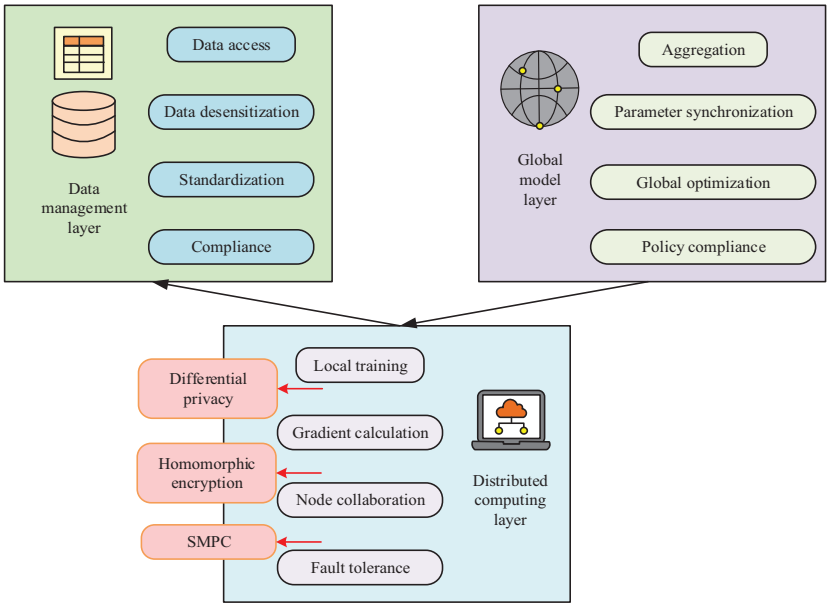

Figure 1 Overall structure of the DML framework for tax big data.

leaked during cross-regional and cross-institutional collaborative modeling processes. In practical business environments, model training must have good robustness and scalability to cope with complex situations such as dynamic node joining or exiting, network condition fluctuations, and uneven data distribution [16]. Based on the above requirements, a DML framework for tax big data is proposed, as shown in Figure 1.

In Figure 1, the DML framework mainly consists of three core levels. The data management layer is responsible for the access, anonymization, and standardization of tax data from different sources, ensuring semantic consistency and data compliance. The distributed computing layer consists of multiple tax nodes, each of which independently stores and trains local data, and uses DML algorithm to calculate local gradients. The global model layer completes gradient aggregation and parameter synchronization through a central aggregation server or decentralized communication mechanism to obtain the global optimal model. Based on the "local training-global aggregation" mode, the computing tasks of multi-source heterogeneous data are decomposed into different tax nodes. While achieving parallel and efficient computing, differential privacy, homomorphic encryption and other technologies are used to ensure data security and compliance, providing solid technical support for

applications such as tax risk identification, invoice anomaly detection, and taxpayer credit evaluation. Based on the DML framework, there is a i-th node with a local dataset of $D_i = [(x_j, y_j)]_{j=1}^{n_i}$. $x_j \in \mathbb{R}^d$ represents the input feature vector of the j-th sample, with dimension d. $y_j \in \mathbb{R}^d$ represents the label value corresponding to the j-th sample. n_i represents the number of samples of local data for the i-th node. Therefore, the local loss function of the i-th node is shown in Equation (1).

$$f_i(\theta) = \frac{1}{n_i} \sum_{j=1}^{n_i} \ell(h_\theta(x_i), y_j) \tag{1}$$

In Equation (1), θ represents the model parameter. $h_\theta(x_i)$ represents the prediction function based on parameter θ. $\ell(\bullet)$ represents the loss function, used to measure the difference between the predicted value and the true value. The mathematical expression for the global optimization objective of DML is shown in Equation (2).

$$\min_\theta F(\theta) = \frac{1}{N} \sum_{i=1}^{N} f_i(\theta) + \lambda \|\theta\|^2 \tag{2}$$

In Equation (2), $F(\theta)$ represents the global loss function, which integrates the local losses of all nodes. N represents the total number of tax nodes participating in distributed training. λ represents the regularization coefficient, which is used to control the complexity of the model and prevent over-fitting. $\|\theta\|^2$ represents the L2 norm regularization term. Each node independently calculates the local gradient, as shown in Equation (3).

$$g_i = \nabla f_i(\theta) \tag{3}$$

In Equation (3), g_i represents the gradient calculated by the i-th node based on local data. $\nabla f_i(\theta)$ represents the gradient of the loss function with respect to the parameter θ. The global gradient g can be obtained by uploading the gradients of each node to the aggregation server through the communication mechanism, as shown in Equation (4).

$$g = \frac{1}{N} \sum_{i=1}^{N} g_i \tag{4}$$

According to the global gradient vector, the update rule of parameter θ can be regarded as $(\theta - \eta g) \to \theta$, where η represents the learning rate. This

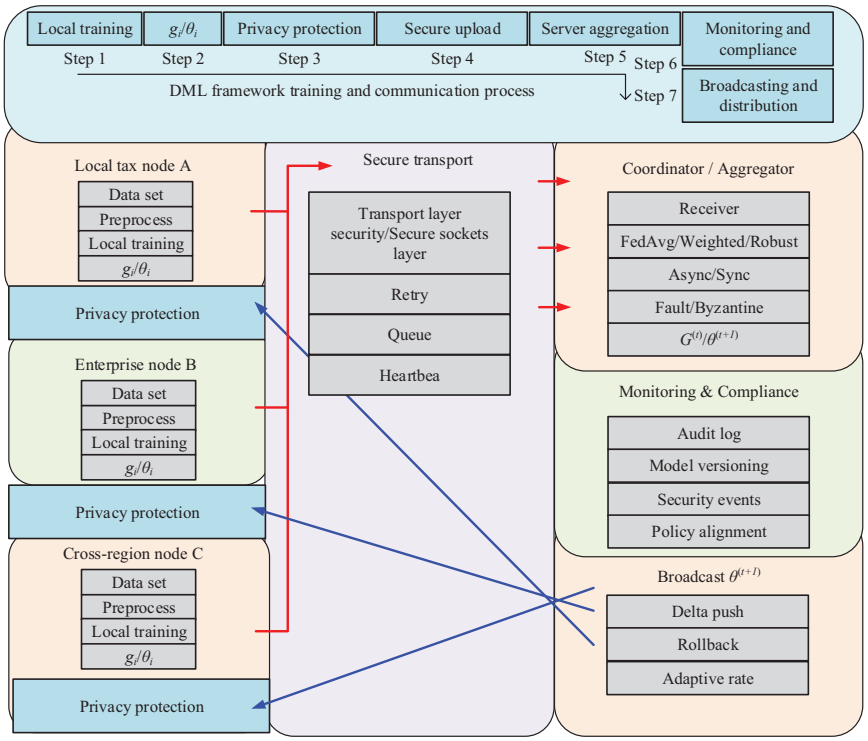

Figure 2 DML framework training and communication process.

process ensures that different nodes can collaborate to optimize the shared model without directly exchanging raw tax data. Based on the above, the training and communication process of DML framework can be shown in Figure 2.

As shown in Figure 2, the training and communication process of DML in tax big data scenarios is divided into four stages: local training, parameter upload, global aggregation, and parameter issuance. During the local training phase, tax nodes such as local tax bureaus and enterprise end systems independently perform model training on the local dataset, calculate the local gradient g_i, and update the parameter θ. During the parameter upload phase, each node encrypts the locally calculated gradient g_i or parameter θ and sends it to the central coordination server. In a distributed environment, there are significant differences in node computing capabilities, and some nodes may experience delays or packet loss. Therefore, the upload process needs to be coordinated with asynchronous/synchronous scheduling strategies. In

the global aggregation stage, the central coordination server aggregates the received local parameter θ_i. During the parameter distribution phase, the updated global model parameter $\theta^{(t+1)}$ is broadcasted back to each node as the initial value for the next round of local training. Through multiple rounds of iterative interaction, the global model ultimately converges under the collaborative action of all distributed nodes. The mathematical expression for parameter aggregation is shown in Equation (5).

$$\theta^{(t+1)} = \sum_{i=1}^{N} \frac{n_i}{\sum_{J=1}^{N} n_j} \theta_i^{(t)} \tag{5}$$

In Equation (5), $\theta_i^{(t)}$ represents the global update parameter of the t-th round. n_i represents the sample size of node i. The DML framework achieves cross-node collaborative optimization in the tax big data scenario through an iterative mode of "local training-global aggregation".

2.2 DML-PP Algorithm Combining Privacy Protection Gradient and Parameter Transmission Mechanism

In the DML framework, although nodes do not directly exchange raw data, the gradients and parameters uploaded and transmitted may also leak privacy. Previous studies have shown that attackers can infer the original sample features through GIA or MRA, and then recover taxpayer information such as invoice amount, transaction time, and identity features [17, 18]. Therefore, a distributed mechanism that relies solely on data without leaving the domain still poses security risks. To address this challenge, a gradient and parameter transmission mechanism for privacy protection is further designed based on the DML framework, and a DML-PP algorithm for tax big data privacy protection is proposed. The overall process of the DML-PP algorithm is shown in Figure 3.

As shown in Figure 3, the DML-PP algorithm first completes model training on the local dataset and calculates the gradient g_i. Before uploading to the aggregation server, the gradient needs to go through a privacy protection processing module. This module includes three key mechanisms: differential privacy perturbation, homomorphic encryption, and SMPC. Differential privacy prevents attackers from recovering individual sample features through gradient inversion by injecting noise into the gradient [19]. Homomorphic encryption ensures the confidentiality of gradients during transmission, allowing aggregation servers to complete addition operations

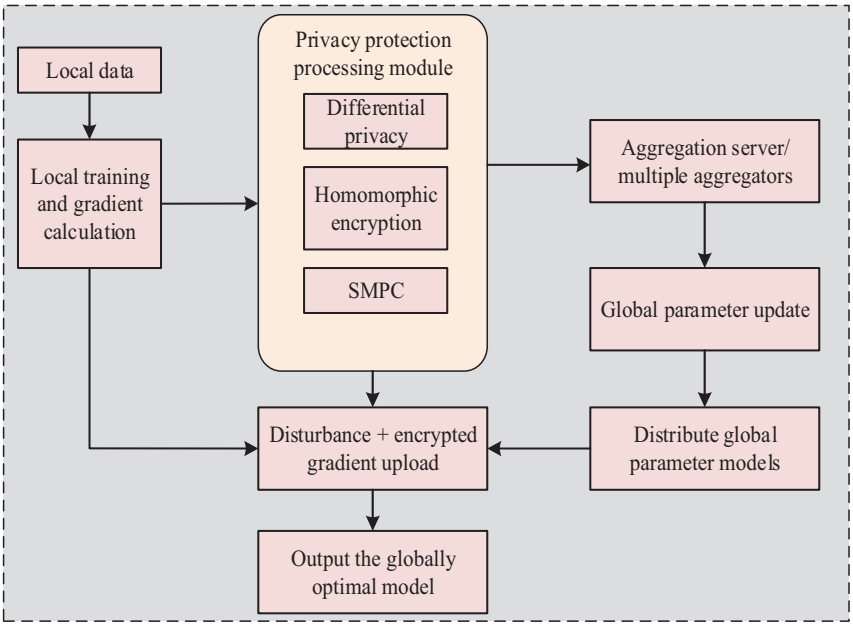

Figure 3 DML-PP algorithm implementation framework.

without decryption [20]. SMPC further eliminates the trust risk of a single aggregation node, enabling cross institutional gradient aggregation to occur in a secure and controllable environment [21]. Finally, the aggregation server summarizes the protected gradients, obtains global model parameters, and distributes them to each node to complete iterative updates. The mathematical expression of the perturbed gradient g_i locally calculated by the i-th node is shown in Equation (6).

$$\tilde{g}_i = g_i + \lambda(0, \sigma^2) \tag{6}$$

In Equation (6), $\tilde{g}_i$ represents the upload gradient after adding noise. $\lambda(0, \sigma^2)$ represents a Gaussian noise distribution with a mean of 0 and a variance of σ^2. The core idea of differential privacy is to add random noise before gradient upload, so that the influence of a single sample on the final upload result is masked, thereby ensuring that attackers cannot accurately infer the information of a specific taxpayer sample even if they obtain the uploaded gradient [22, 23]. The specific implementation process is shown in Figure 4.

In Figure 4, firstly, each node trains the model based on local data and calculates the gradient g_i. Secondly, the system uses a random mechanism to

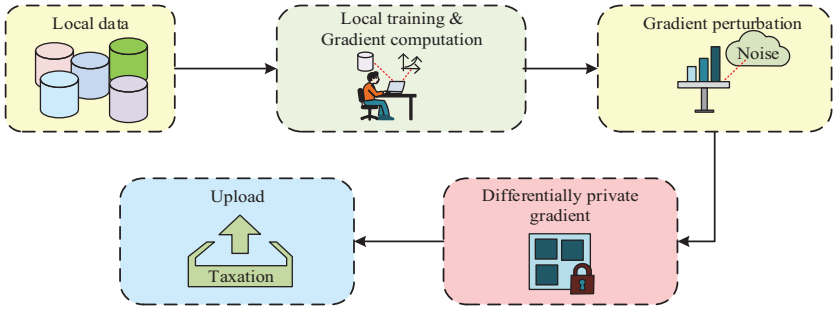

Figure 4 Local gradient upload process under differential privacy perturbations.

generate noise terms from Gaussian distribution and superimposes them with the gradient to obtain the perturbed upload gradient $\tilde{g}_i$. Finally, the nodes upload $\tilde{g}_i$ to the aggregation server. Based on this mechanism, even if the gradient upload process is intercepted, attackers cannot recover the original sample features, thus ensuring the privacy of individual taxpayer data. After completing the differential privacy perturbation, the node homomorphic encrypts the perturbed gradient $\tilde{g}_i$. Assuming that the encryption function is $\mathrm{Enc}(\bullet)$, there is a mathematical expression in Equation (7).

$$c_i = \mathrm{Enc}(\tilde{g}_i) \tag{7}$$

In Equation (7), c_i represents the ciphertext gradient uploaded by node i. At this point, the central aggregation server can directly perform addition operations in the ciphertext space without decryption, and obtain relevant mathematical equations through the decryption function $\mathrm{Dec}(\bullet)$, as shown in Equation (8).

$$\begin{cases} C = \sum_{i=1}^{N} c_i \\ \mathrm{Dec}(C) = \sum_{i=1}^{N} \tilde{g}_i \end{cases} \tag{8}$$

In Equation (8), C represents the ciphertext collection calculated by the aggregation server. $\sum_{i=1}^{N} \tilde{g}_i$ represents the perturbation set of gradients uploaded by all nodes. The gradient aggregation and decryption flowchart under homomorphic encryption is shown in Figure 5.

In Figure 5, after completing gradient perturbation locally, each node encrypts it and generates ciphertext c_i, which is then uploaded to the aggregation server. The server can directly add all ciphertexts without decrypting

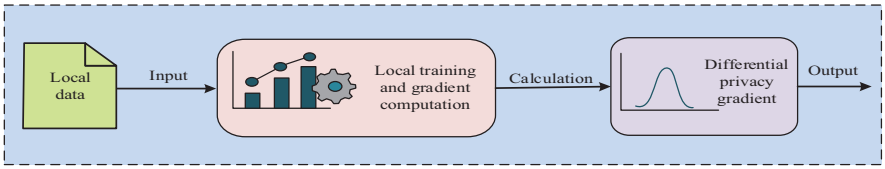

Figure 5 Gradient aggregation and decryption flowchart.

to obtain the aggregation result C. After the aggregation phase is completed, the entity with the decryption key performs the decryption operation to recover the perturbed global gradient vector. Considering that in cross-institutional and cross-regional tax big data scenarios, the participating nodes usually belong to different trust domains [24]. If relying solely on a central aggregation server, it may pose a single point trust risk [25]. Therefore, the study further introduces the SMPC protocol to collectively complete global gradient aggregation without exposing their respective inputs. Each node i divides the local gradient g_i into M random shares. Each participating node only holds a partial share, and no single party can restore the original gradient. Ultimately, during protocol execution, the global perturbation gradient is obtained by aggregating all shares. The specific calculation is shown in Equation (9).

$$\begin{cases} g_i = \sum_{i=1}^{M} s_{i,k} \\ g = \frac{1}{N} \sum_{i=1}^{N} g_i \end{cases} \tag{9}$$

In Equation (9), $s_{i,k}$ represents the k-th secret share generated by node i. The SMPC protocol can ensure the correctness of gradient aggregation and avoid single point trust dependencies, thereby increasing the robustness and verifiability of the system [26]. The aggregation process is shown in Figure 6.

In Figure 6, firstly, node i randomly splits the local gradient g_i into several shares $\{s_{i,k}\}$ and sends them to multiple different aggregation parties. Each aggregator can only see local information and cannot independently infer the specific value of g_i. Finally, only when all aggregation parties cooperate can the shares be combined and the global aggregation be completed, resulting in the perturbed global gradient vector. This mechanism not only avoids excessive trust in a single central server, but also enhances the credibility and security of cross institutional collaboration. In summary, the proposed DML-PP algorithm includes triple protection, which can achieve a balance

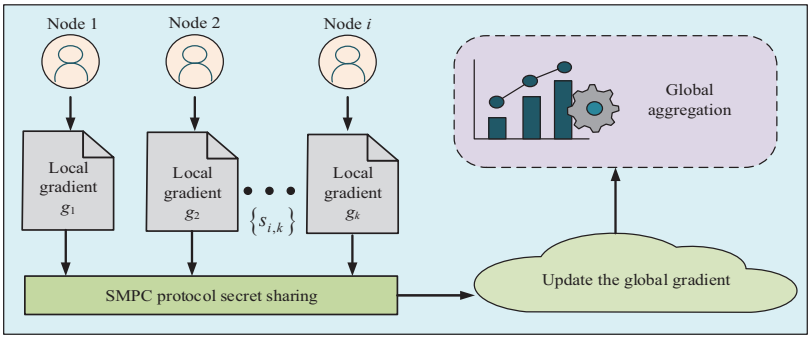

Figure 6 SMPC-based secret sharing and global aggregation process.

between system performance and security. In addition, the triple mechanism can not only meet the compliance requirements in the tax big data, but also provide solid technical support for tasks such as invoice anomaly detection, credit evaluation, and anti-tax avoidance modeling.

3 Results and Analysis

Firstly, the experimental environment and dataset construction are introduced, including the basic characteristics and processing methods of real tax data and simulated data. Secondly, the performance of the proposed algorithm is compared with other algorithms. Finally, the testing is conducted from the perspectives of security and robustness, covering complex scenarios such as poisoning attacks, node failures, network packet loss, and policy compliance.

3.1 Experimental Setup

To verify the effectiveness and feasibility of the proposed DML-PP algorithm, systematic experiments are conducted in real data and simulated experimental environments. The experiment is conducted on a distributed computing server cluster. The hardware and software configurations of each node are shown in Table 1.

The experimental dataset consists of a real tax dataset and a synthetic simulation dataset. The real tax dataset is the invoice circulation and tax declaration data anonymized by a provincial tax bureau, with 1.2 million records and fields covering taxpayer number, industry category, invoice amount, declaration period, etc. To protect data privacy, all sensitive fields undergo desensitization and encryption processing. The synthetic simulation

Table 1 Experimental environment configuration

Level	Configuration Details	Specific Parameters
Hardware	CPU	Intel Xeon Gold 6330 (28 cores, 2.0 GHz)
	Memory	128 GB DDR4
	GPU	NVIDIA A100 (40 GB memory)
Software	Operating system	Ubuntu 22.04 LTS
	Distributed communication framework	Message passing interface
	Deep learning framework	TensorFlow 2.10, PyTorch 2.0
	Privacy protection tools	PySyft, Microsoft SEAL

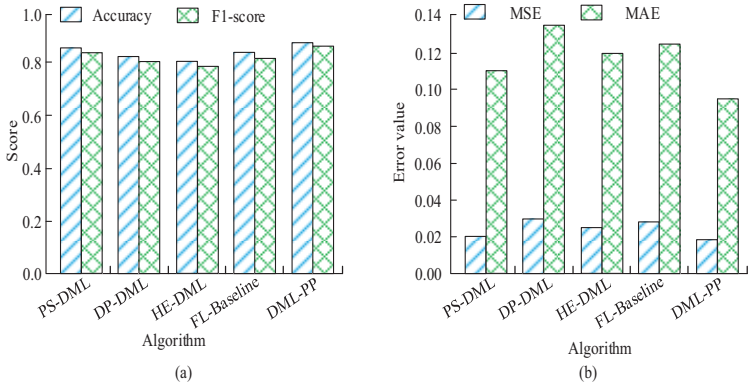

Figure 7 Comparison of prediction results from different algorithms (a) Comparison of classification task accuracy and F1-score; (b) Comparison of regression task errors.

dataset is a simulation dataset generated based on a real tax dataset, with 5 million samples and features including transaction amount, frequency, enterprise size, declaration anomaly rate, etc. It simulates the distributed modeling process of cross-regional tax nodes.

3.2 Comparison of Algorithm Performance

To comprehensively evaluate the effectiveness of the DML-PP algorithm, the study introduces the DML based on Differential Privacy (DP-DML) [27], Plain DML (PS-DML) without introducing any privacy protection mechanism [28], DML based on Homomorphic Encryption (HE-DML) [29], and Federated Learning Baseline (FL-Baseline) [30–32] for performance comparison. The prediction results of different algorithms are shown in Figure 7.

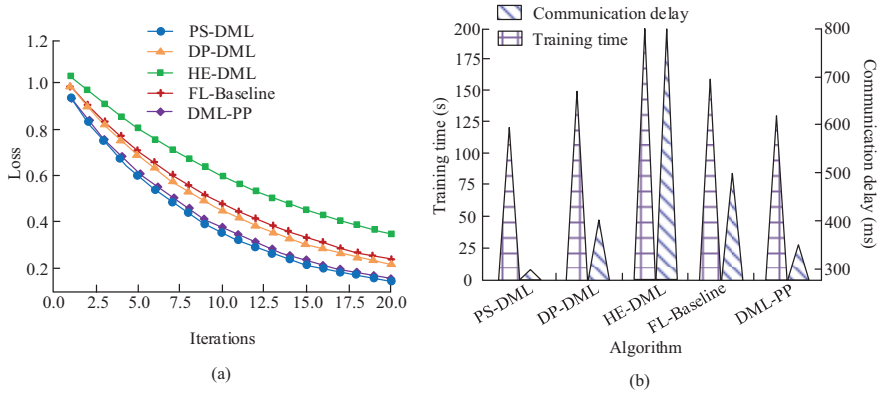

Figure 8 Comparison of computational efficiency of different algorithms (a) Convergence curve comparison; (b) Comparison of average training time and communication delay.

From Figure 7(a), the accuracy of DML-PP in classification tasks reached 0.87, which was 2.35%, 6.10%, 8.75%, and 4.82% higher than that of PS-DML, DP-DML, HE-DML, and FL-Baseline, respectively. In terms of F1-score, DML-PP also performed the best, reaching 0.86. This indicates that the DML-PP algorithm can still maintain superior performance under privacy protection conditions. Combined the results of Mean Square Error (MSE) and Mean Absolute Error (MAE) in Figure 7(b), the prediction error of DML-PP in regression tasks was significantly lower than that of other algorithms. Compared with PS-DML, DP-DML, HE-DML, and FL-Baseline, the MSE of DML-PP decreased by 10.00%, 40.00%, 28.00%, and 28.00%, respectively, while the MAE decreased by 22.03%. DML-PP not only performs superior in classification tasks, but also demonstrates stronger generalization ability and stability in regression tasks. The computational efficiency of different algorithms is shown in Figure 8.

From Figure 8(a), the convergence speed of DML-PP was faster than that of DP-DML, HE-DML, and FL-Baseline, only slightly lower than that of PS-DML. This indicates that during multiple rounds of iterations, DML-PP can approach the optimal solution faster, mainly due to its efficient design in gradient perturbation and encryption calculation, which reduces invalid updates. According to Figure 8(b), the training time of DML-PP was 130.24 seconds, which was 34.93%, 13.38%, and 18.80% shorter than that of HE-DML, DP-DML, and FL-Baseline, respectively. The advantage of training time mainly comes from the lightweight perturbation strategy in the privacy protection mechanism, which significantly reduces the computational cost

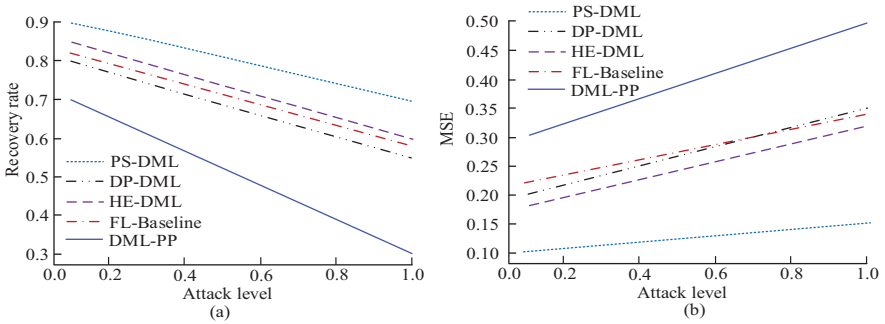

Figure 9 Comparison of privacy protection effects of different algorithms (a) Comparison of GIA recovery rates; (b) Comparison of data reconstruction errors.

of encryption and decryption while ensuring security. Although the training time of DML-PP has slightly increased compared to PS-DML, considering its additional privacy protection measures, this gap is still within an acceptable range, indicating that DML-PP has achieved a good balance between security and computational efficiency. In addition, the communication delay of DML-PP was 350.18 ms, significantly lower than that of HE-DML, DP-DML, and FL-Baseline. This is thanks to its optimized gradient compression and aggregation mechanism, which effectively reduces the scale of data transmission across nodes. Compared with PS-DML, although the communication delay of DML-PP increased by 16.68%, the difference was mainly due to additional privacy perturbation operations. Overall, the delay level is still relatively low and will not have a significant impact on the synchronization efficiency in the distributed learning process. Therefore, DML-PP achieves privacy protection and performance optimization while improving convergence speed and maintaining low delay, demonstrating strong practical application potential. The privacy protection effects of different algorithms are shown in Figure 9.

From Figure 9(a), PS-DML maintained a high recovery rate as the attack intensity gradually increased. Without privacy protection, attackers could recover training data with high accuracy. The recovery rates of DP-DML and HE-DML both decreased, but the lowest remained at 0.55–0.60. In contrast, the recovery rate of DML-PP was only 0.30 when the attack level was 1.0, which was 45.45%, 50.00%, 48.28%, and 57.14% lower than that of DP-DML, HE-DML, FL-Baseline, and PS-DML, respectively. This indicates that DML-PP performs the best in resisting GIA and can reduce the exposure risk of sensitive data. In Figure 9(b), the reconstruction error of PS-DML was only 0.10–0.15, indicating that attackers can reconstruct the original

Table 2 Performance comparison of different algorithms

Algorithm	Epochs	Communication Overhead (MB/Epoch)	Node Fault Tolerance Rate (%)	Total Running Time (s)
PS-DML	12	25.37	95.12	1201.46
DP-DML	15	28.64	93.46	1452.73
HE-DML	20	80.42	92.73	1983.58
FL-Baseline	18	40.27	94.18	1607.35
DML-PP	13	32.41	96.38	1301.82

data with high accuracy and privacy protection is almost ineffective. The reconstruction errors of DP-DML and HE-DML were improved, with the highest being 0.35 and 0.32, respectively. In contrast, the reconstruction error of DML-PP reached 0.50 when the attack level was 1.0. This indicates that DML-PP can significantly increase the reconstruction difficulty for attackers and effectively protect the privacy information of the original samples. The communication overhead and operational efficiency of different algorithms are shown in Table 2.

From Table 2, DML-PP achieved convergence in the 13th epoch, and its convergence efficiency was significantly better than that of the method using a single privacy protection mechanism. The reason for this result is that DML-PP effectively suppresses the accuracy loss caused by gradient perturbations through differential privacy and lightweight encryption mechanisms, and avoids the redundant computational overhead generated by homomorphic encryption in high round epochs. The average communication overhead of DML-PP was 32.41 MB/round, which was 59.67% and 19.52% lower than that of HE-DML and FL-Baseline, respectively, and only slightly higher than that of DP-DML. This indicates that DML-PP achieves a good balance between privacy protection and communication overhead. The reason for this advantage is that DML-PP combines compression and perturbation strategy in the gradient upload process, which not only reduces the transmission data size, but also preserves necessary model update information, thus achieving bandwidth optimization without sacrificing too much accuracy. In the case of simulating 10% node disconnection, the fault tolerance of DML-PP reached 96.38%, which was higher than that of the other four algorithms. This indicates that DML-PP has stronger robustness in heterogeneous and unstable network environments. Due to its redundant updates and local aggregation mechanisms, DML-PP can maintain stable updates of the global model even in the absence of some nodes. Based on the overall running time, DML-PP not only achieves a balance between convergence speed and communication

overhead, but also shows significant advantages in fault tolerance. This further demonstrates that the method can balance efficiency and stability while ensuring privacy protection, demonstrating strong engineering feasibility and practical application value.

3.3 Security and Robustness Analysis

To evaluate the adaptability of DML-PP in complex tax scenarios, the study first compares the robustness of different algorithms in the malicious node poisoning scenario, as shown in Figure 10.

As shown in Figure 10(a), as the poisoning ratio increased, the model accuracy of PS-DML decreased from 0.87 to 0.30, with a decrease of 65.52%. FL-Baseline decreased by 59.30%, while the accuracy of DP-DML and HE-DML remained at 0.40 and 0.38, respectively, at a 30% poisoning ratio, with a decrease of 52.94% and 54.76% from the initial values, respectively. In contrast, the accuracy of DML-PP only decreased by 25.29%, which was significantly better than that of other methods. This indicates that DML-PP has stronger resistance to poisoning. Combining Figure 10(b) can further illustrate the convergence stability of the DML-PP algorithm. To test the ability of DML-PP to cope with node disconnections and failures in distributed systems, the study sets node failure rates of 0%, 5%, 10%, and 20% respectively, and compares the performance of different algorithms in global model accuracy and fault tolerance. The experimental results are shown in Figure 11.

As shown in Figure 11(a), DML-PP still exhibited strong resistance to failure at different node failure rates, with strong robustness against failure.

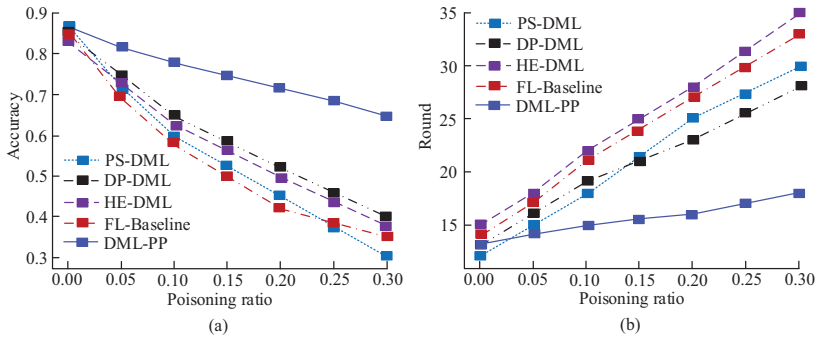

Figure 10 Robustness comparison of poisoning attacks (a) Changes in model accuracy under poisoning ratios; (b) Convergence round changes under poisoning ratio.

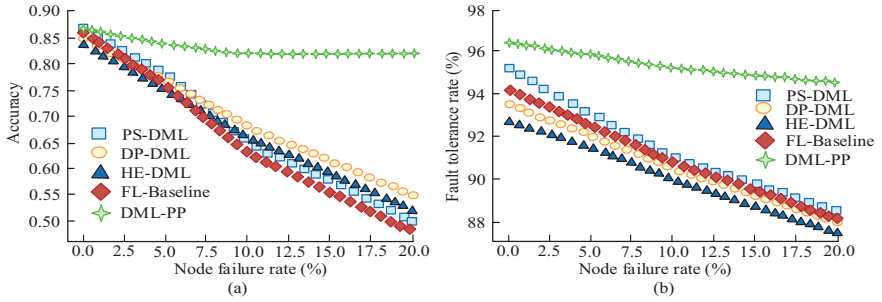

Figure 11 Comparison of node failure robustness (a) Changes in model accuracy under node failure rates; (b) Convergence round changes under poisoning ratio.

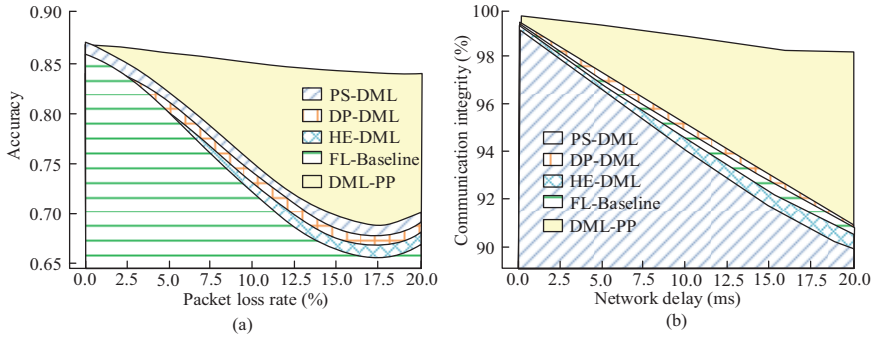

Figure 12 Comparison of communication security and stability (a) Changes in model accuracy under different packet loss rates; (b) Convergence round changes under poisoning ratio.

Figure 11(b) compares the fault tolerance under different failure rates, which further demonstrates the superiority of DML-PP. Compared with other algorithms, the DML-PP algorithm improved its fault tolerance by 6.78%, 7.39%, 8.00%, and 7.14% at a node failure rate of 20%. This indicates that DML-PP has superior robustness. The communication security of different algorithms is shown in Figure 12.

From Figure 12(a), at a packet loss rate of 20%, the accuracy of the DML-PP model was 0.84, which was 20.00%, 21.74%, 23.53%, and 25.37% higher than that of PS-DML, DP-DML, HE-DML, and FL-Baseline, respectively. The accuracy of DML-PP only decreased from 0.87 to 0.84, with a decrease of 3.45%. Other algorithms had a more significant decrease. This indicates that DML-PP has better stability in high packet loss environments. From Figure 12(b), at a network delay of 200 ms, the communication integrity

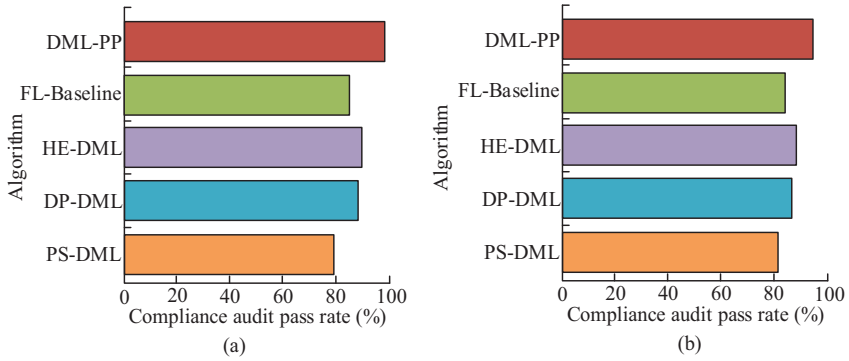

Figure 13 Comparison of policy compliance and verifiability (a) Comparison of compliance audit pass rate; (b) Comparison of system security incident detection rates.

of DML-PP was 98.10%, significantly higher than that of the other four algorithms, and the decrease was even lower. This indicates that DML-PP can maintain highly stable communication security even with increased delay. This study further simulates audit processes and security event detection experiments to evaluate the policy compliance and verifiability of the DML-PP algorithm in the tax big data scenario, as shown in Figure 13.

According to Figure 13(a), the compliance audit pass rate of DML-PP was 98.70%, which was 24.29%, 11.53%, 9.42%, and 15.62% higher than that of PS-DML, DP-DML, HE-DML, and FL-Baseline, respectively. Based on Figure 13(b), DML-PP not only had advantages in privacy protection and compliance, but also outperformed other algorithms in system verifiability. Compared with other algorithms, the detection rates of DML-PP increased by 15.92%, 8.80%, 6.97%, and 12.95%, respectively. This indicates that DML-PP not only meets policy compliance requirements, but also has superior security audit capabilities.

4 Discussion

The experimental results showed that DML-PP achieved higher accuracy and lower error in both classification and regression tasks, with classification accuracy and F1-score reaching 0.87 and 0.86, respectively. The MSE and MAE values were lower than those of the comparison algorithm. This is consistent with the conclusion proposed by Tan et al. in reference [10] that intelligent tax audit systems can improve risk identification efficiency. The reason is that the proposed algorithm considers both gradient perturbation and

homomorphic encryption in the privacy protection transmission mechanism, allowing the model to obtain a relatively stable optimization process while ensuring data security. However, there are differences from some research results. For example, Liu et al. found that the privacy protection mechanism reduces the convergence speed and prediction accuracy of the model in the distributed edge computing scenario in reference [11]. DML-PP achieves a better balance between efficiency and accuracy. This may be due to the structured features of tax data and optimized aggregation strategies, which effectively control communication overhead within the overall framework.

In terms of privacy protection effectiveness, DML-PP performed the best in resisting GIA and MRA, with the lowest attack recovery rate of only 0.30 and a reconstruction error of 0.50. This result is highly consistent with the "multiple protection" approach emphasized in the five-dimensional privacy protection framework proposed by Padmanaban in reference [12]. The reason is that the research combines differential privacy, homomorphic encryption, and SMPC protocol organically, avoiding the vulnerability of a single mechanism. However, compared with the AI penetration supervision model proposed by Zhu et al. in financial big data supervision in reference [13], the research results are different. Zhu et al. found that the privacy protection in cross-border data flow can reduce the sensitivity of real-time monitoring. This difference may be due to the higher cross-border liquidity and complexity of financial data compared to tax invoice and declaration data, resulting in a decrease in the performance of protection mechanisms in high time scenarios, while tax data is relatively stable and easier to balance security and accuracy.

Robustness analysis shows that DML-PP maintains high accuracy and fault tolerance in complex scenarios such as poisoning attacks, node failures, and network packet loss, which is significantly better than comparison algorithms. This may be because DML-PP eliminates single point of trust dependencies through the SMPC protocol, improving the robustness of the system in untrusted environments. In terms of policy compliance and security audits, the compliance audit pass rate of DML-PP reached 98.70%, indicating that it not only meets compliance requirements, but also has strong verifiability. This is similar to the inter regional value-added tax compliance optimization design results proposed by Chen et al. based on e-commerce big data in reference [2]. However, some scholars have put forward different views. For example, in reference [7], Ariyibi et al. pointed out in the tax anti-fraud framework that overly strong privacy protection mechanisms may weaken the interpretability and transparency of the model. In contrast, the research emphasizes more on "privacy first". Therefore, in practical

promotion, a further balance between interpretability and privacy is still necessary.

Although the DML-PP algorithm has shown strong advantages in prediction accuracy, privacy protection, and robustness, there are still some limitations. Firstly, the experiment mainly employs real data and synthetic simulation data from provincial tax bureaus for verification. The data types are relatively concentrated, which may limit the adaptability of the algorithm in the global tax environment. Secondly, while ensuring privacy, DML-PP still has slightly higher communication overhead than some single mechanism models, which may bring additional burden in large-scale high-frequency interaction environments. In addition, the research has not yet delved into the shortcomings on model interpretability, which is of great significance for tax regulation and judicial auditing. Therefore, future work can be carried out in the following aspects. Firstly, the experimental data sources can be further expanded to include cross-border transactions, e-commerce, and international tax compliance scenarios to enhance generality and applicability. Secondly, more efficient encryption and aggregation protocols can be explored, such as homomorphic encryption methods based on quantum security, to further reduce communication and computational overhead. It is also possible to consider introducing an interpretability enhancement mechanism that combines privacy protection with interpretability, ensuring that the algorithm can provide clear and transparent decision-making basis for regulatory authorities while protecting taxpayers' privacy.

5 Conclusion

This study addresses the prominent challenge of balancing privacy protection and computational efficiency in cross-regional collaborative modeling of tax big data. A privacy-preserving algorithm framework DML-PP was constructed. The results demonstrate that the proposed algorithm outperforms traditional methods in both classification and regression tasks, improving classification accuracy by 2.35%–8.75%, reducing communication overhead by 59.67%, and achieving a 96.38% fault tolerance in node failure scenarios. Unlike existing single privacy-preserving mechanisms, this study innovatively integrates differential privacy, homomorphic encryption, and SMPC in a multi-layered manner, constructing a privacy-preserving chain encompassing local training, secure transmission, aggregated computation, and global synchronization, enhancing the security and stability of the distributed modeling process. This mechanism not only addresses the shortcomings of existing

literature in balancing privacy protection and performance, but also provides a more feasible technical approach for intelligent tax governance. The study validates that this multi-mechanism collaborative privacy-preserving strategy effectively defends against gradient inversion and model reconstruction attacks without sacrificing model performance, providing strong support for the secure sharing and joint modeling of multi-source tax data. Future research will further expand the application potential of this algorithm in complex scenarios such as cross-border tax coordination and financial risk control, and explore lightweight encryption and interpretability enhancement mechanisms to achieve more efficient and transparent distributed intelligent governance models. Furthermore, the algorithm will integrate policy and technical requirements to build a privacy-preserving and collaborative computing framework adaptable to diverse compliance environments, providing a more versatile and scalable solution for large-scale distributed data analysis.

References

[1] O. Tuyishimire and B. F. Murorunkwere. "Applications of big data analytics in tax compliance monitoring: A case study of Rwanda's value-added tax," CESifo Econ. Stud., vol. 70, no. 4, pp. 578–587, December, 2024, DOI: 10.1093/cesifo/ifae027.

[2] Y. Chen, L. Xiang, and H. Yang. "Interregional value-added tax in the era of e-commerce: Tax policy design based on big data from online retailing," J. Social Comput., vol. 5, no. 1, pp. 46–57, January, 2024, DOI: 10.23919/JSC.2024.0006.

[3] R. Belahouaoui and E. H. Attak. "Digital taxation, artificial intelligence and tax administration 3.0: Improving tax compliance behavior–a systematic literature review using textometry (2016–2023)," Account. Res. J., vol. 37, no. 2, pp. 172–191, March, 2024, DOI: 10.1108/ARJ-12-2023-0372.

[4] M. Hasanvand, M. Nooshyar, E. Moharamkhani, and A. Selyari. "Machine learning methodology for identifying vehicles using image processing," AIA, vol. 1, no. 3, pp. 170–178, April, 2023, DOI: 10.47852/bonviewAIA3202833.

[5] R. A. S. R. Wahab and A. Bakar. "Digital economy tax compliance model in Malaysia using machine learning approach," Sains Malays., vol. 50, no. 7, pp. 2059–2077, July, 2021, DOI: 10.17576/jsm-2021-5007-20.

[6] S. Ullah, R. Luo, T. S. Adebayo, and M. T. Kartal. "Dynamics between environmental taxes and ecological sustainability: Evidence from top-seven green economies by novel quantile approaches," Sustain. Dev., vol. 31, no. 2, pp. 825–839, February, 2023, DOI: 10.1002/sd.2423.

[7] K. O. Ariyibi, O. F. Bello, T. F. Ekundayo, and O. Ishola. "Leveraging artificial intelligence for enhanced tax fraud detection in modern fiscal systems," GSC Adv. Res. Rev., vol. 21, no. 2, pp. 129–137, February, 2024, DOI: 10.30574/gscarr.2024.21.2.0415.

[8] A. Atadoga, U. J. Umoga, O. A. Lottu, and E. O. Sodiya. "Evaluating the impact of cloud computing on accounting firms: A review of efficiency, scalability, and data security," Global J. Eng. Technol. Adv., vol. 18, no. 2, pp. 65–74, February, 2024, DOI: 10.30574/gjeta.2024.18.2.0027.

[9] A. A. Elamer, M. Boulhaga, and B. A. Ibrahim. "Corporate tax avoidance and firm value: The moderating role of environmental, social, and governance (ESG) ratings," Bus. Strategy Environ., vol. 33, no. 7, pp. 7446–7461, July, 2024, DOI: 10.1002/bse.3881.

[10] Y. Tan, W. Zheng, J. Cao, and B. Jiang. "Intelligent tax systems: Automating tax audits and improving revenue efficiency," Open J. Account., vol. 14, no. 3, pp. 156–169, March, 2025, DOI: 10.4236/oj acct.2025.143009.

[11] S. Liu, C. Zheng, Y. Huang, and T. Q. Quek. "Distributed reinforcement learning for privacy-preserving dynamic edge caching," IEEE J. Sel. Areas Commun., vol. 40, no. 3, pp. 749–760, March, 2022, DOI: 10.1109/JSAC.2022.3142348.

[12] H. Padmanaban. "Privacy-preserving architectures for AI/ML applications: Methods, balances, and illustrations," J. Artif. Intell. Gen. Sci., vol. 3, no. 1, pp. 235–245, January, 2024, DOI: 10.60087/jaigs.v3i1.117.

[13] Y. Zhu, K. Yu, M. Wei, Y. Pu, and Z. Wang. "AI-enhanced administrative prosecutorial supervision in financial big data: New concepts and functions for the digital era," J. Adv. Comput. Syst., vol. 4, no. 5, pp. 10–26, May, 2024, DOI: 10.69987/JACS.2024.40502.

[14] M. Li, F. Wang, X. Jia, W. Li, T. Li, and G. Rui. "Multi-source data fusion for economic data analysis," Neural Comput. Appl., vol. 33, no. 10, pp. 4729–4739, May, 2021, DOI: 10.1007/s00521-020-05531-0.

[15] Y. Tong and R. Zhang. "Investigating the multiple mechanisms of tourism economy affecting sustainable urban development of Chinese cities: Based on multi-source data," Environ. Dev. Sustain., vol. 26, no. 1, pp. 1781–1808, January, 2024, DOI: 10.1007/s10668-022-027 85-7.

[16] L. Shen, Y. Sun, Z. Yu, L. Ding, X. Tian, and D. Tao. "On efficient training of large-scale deep learning models," ACM Comput. Surv., vol. 57, no. 3, pp. 1–36, November, 2024, DOI: 10.1145/3700439.

[17] D. Usynin, D. Rueckert, and G. Kaissis. "Beyond gradients: Exploiting adversarial priors in model inversion attacks," ACM Trans. Privacy Secur., vol. 26, no. 3, pp. 1–30, June, 2023, DOI: 10.1145/3592800.

[18] R. Yang, J. Ma, J. Zhang, S. Kumari, S. Kumar, and J. J. Rodrigues. "Practical feature inference attack in vertical federated learning during prediction in artificial Internet of Things," IEEE Internet Things J., vol. 11, no. 1, pp. 5–16, May, 2023, DOI: 10.1109/JIOT.2023.3275161.

[19] Y. Zhao and J. Chen. "A survey on differential privacy for unstructured data content," ACM Comput. Surv., vol. 54, no. 10s, pp. 1–28, September, 2022, DOI: 10.1145/3490237.

[20] K. Munjal and R. Bhatia. "A systematic review of homomorphic encryption and its contributions in healthcare industry," Complex Intell. Syst., vol. 9, no. 4, pp. 3759–3786, August, 2023, DOI: 10.1007/s40747-022-00756-z.

[21] V. Sucasas, A. Aly, G. Mantas, J. Rodriguez, and N. Aaraj. "Secure multi-party computation-based privacy-preserving authentication for smart cities," IEEE Trans. Cloud Comput., vol. 11, no. 4, pp. 3555–3572, July, 2023, DOI: 10.1109/TCC.2023.3294621.

[22] B. Jia, X. Zhang, J. Liu, Y. Zhang, K. Huang, and Y. Liang. "Blockchain-enabled federated learning data protection aggregation scheme with differential privacy and homomorphic encryption in IIoT," IEEE Trans. Ind. Informat., vol. 18, no. 6, pp. 4049–4058, June, 2021, DOI: 10.1109/TII.2021.3085960.

[23] J. Dong, A. Roth, and W. J. Su. "Gaussian differential privacy," J. R. Stat. Soc. Ser. B-Stat. Methodol., vol. 84, no. 1, pp. 3–37, February, 2022, DOI: 10.1111/rssb.12454.

[24] Y. Liu and L. Buckingham. "Academic research network management: Sociocultural perspectives from languages other than English," J. Lang. Identity Educ., vol. 24, no. 4, pp. 841–857, 2025, DOI: 10.1080/15348458.2023.2196629.

[25] Y. Zheng, S. Lai, Y. Liu, X. Yuan, X. Yi, and C. Wang. "Aggregation service for federated learning: An efficient, secure, and more resilient realization," IEEE Trans. Dependable Secure Comput., vol. 20, no. 2, pp. 988–1001, January, 2022, DOI: 10.1109/TDSC.2022.3146448.

[26] M. Ahmad, S. Habib, and F. Tariq. "Enhancing model robustness in federated learning: A systematic literature review of Byzantine-resilient

aggregation methods," VFAST Trans. Softw. Eng., vol. 13, no. 2, pp. 196–227, 2025, DOI: 10.21015/vtse.v13i2.2163.
[27] J. So, B. Güler, and A. S. Avestimehr. "CodedPrivateML: A fast and privacy-preserving framework for distributed machine learning," IEEE J. Sel. Areas Inf. Theory, vol. 2, no. 1, pp. 441–451, March, 2021, DOI: 10.1109/JSAIT.2021.3053220.
[28] J. Wang, A. Pal, Q. Yang, K. Kant, K. Zhu, and S. Guo. "Collaborative machine learning: Schemes, robustness, and privacy," IEEE Trans. Neural Netw. Learn. Syst., vol. 34, no. 12, pp. 9625–9642, December, 2022, DOI: 10.1109/TNNLS.2022.3169347.
[29] J. Chen, K. Li, and S. Y. Philip. "Privacy-preserving deep learning model for decentralized VANETs using fully homomorphic encryption and blockchain," IEEE Trans. Intell. Transp. Syst., vol. 23, no. 8, pp. 11633–11642, August, 2022, DOI: 10.1109/TITS.2021.3105682.
[30] R. Wang, H. Qiu, H. Gao, C. Li, Z. Y. Dong, and J. Liu. "Adaptive horizontal federated learning-based demand response baseline load estimation," IEEE Trans. Smart Grid, vol. 15, no. 2, pp. 1659–1669, September, 2023, DOI: 10.1109/TSG.2023.3318418.
[31] K. Somsuk, "Enhanced Algorithm for Recovering RSA Plaintext when Two Modulus Values Share At least One Common Prime Factor", JCSANDM, vol. 14, no. 02, pp. 433–456, Jun. 2025.
[32] M. Gao, Z. Zhang, L. Cui, S. Feng, J. Liu, and Y. Jiang, "Temporal and Topological Enhanced Graph Neural Networks for Traffic Anomaly Detection", JCSANDM, vol. 14, no. 02, pp. 457–474, Jun. 2025.

Biographies

Ying Fang obtained her Bachelor's Degree in Financial Management from Heilongjiang University of Science and Technology in 2006. She obtained

her Master's Degree in Business Administration from Yanshan University in 2013. Presently, she is working as an associate professor in the Department of Economics, Qinhuangdao Vocational and Technical College. Her areas of interest are financial management, enterprise management, and vocational education.

Ling Pu obtained her Master of Economics from Nankai University in 2007. Presently, she is working as an associate professor in the Department of Economics, Qinhuangdao Vocational and Technical College. Her areas of interest are economics, financial management, and financial digitization.

Ning Zhang obtained her Master of Vocational and Technical Education from Hebei Normal University of Science and Technology in 2015. Presently, she is working as an associate professor in the Department of Economics, Qinhuangdao Vocational and Technical College. Her areas of interest are accounting, education, and financial digitization.

Jun Liang obtained his Bachelor of Management from Hebei University in 2006. Presently, he is working as an associate professor and Director of Teaching in the Department of Economics, Qinhuangdao Vocational and Technical College. His areas of interest are financial management, management accounting, and financial digitization.

Shaojuan Ouyang obtained her Master of Economics from Tianjin University of Finance and Economics in 2012. Presently, she is working as aprofess or in the Department of Economics, Qinhuangdao Vocational and Technical College. Her areas of interest are accounting, financial management, and financial digitization.