Scholar Publishing Group
International Journal of Multimedia Computing
https://doi.org/10.38007/IJMC.2026.070102
ISSN 2789-7168 Vol. 7, Issue 1:9-18

Research on Privacy-Preserving AI Model Training and Validation Methods Based on Federated Learning

Mingjie Chen

Software and Societal Systems Department, School of Computer Science, Carnegie Mellon University, Pittsburgh 15213

Keywords: Federated Learning Privacy Protection; Differential Privacy Adaptive Aggregation

Abstract: Federated learning, as a distributed collaborative modeling framework, achieves joint learning under the premise of protecting data privacy by completing training locally and only sharing model updates. However, in practical applications, there are still risks such as gradient leakage and poisoning attacks. To enhance privacy protection and model robustness, this study proposes a protocol that combines differential privacy, identity-based signatures, ordered encryption and correlation detection to effectively identify and eliminate abnormal gradients. On this basis, an adaptive aggregation method based on local differential privacy is designed. Enable users to flexibly set privacy budgets and maintain the high precision of the global model by leveraging security aggregations. Theoretical analysis and experimental verification show that this method can not only effectively resist malicious attacks, but also improve the convergence and prediction performance of the model while ensuring privacy. It further demonstrates its application value in the secure training of sensitive data and disease prediction in the prototype system of the digital medical scenario.

1. Introduction

Under the centralized learning model, researchers often need to uniformly transfer the raw data scattered across different institutions or terminals to a central node for training. Although this approach can bring direct advantages in model accuracy and optimization efficiency, due to the involvement of highly sensitive information such as medical records, financial data, and travel trajectories, it is very likely to cause privacy leakage and data abuse. At the same time, it will also lead to cross-institutional data isolation due to institutional and technological barriers, thereby restricting the promotion and application of artificial intelligence technology in key fields. To break through this bottleneck, a training paradigm centered on distributed collaboration has gradually emerged. This model allows each participant to independently complete model updates in the local environment, only uploading parameters or weights to the coordination end for aggregation. This enables multi-party joint modeling without disclosing the original data and effectively alleviates the problem of data silos.

Although this collaborative framework provides a new solution for privacy protection in concept,

as its application continues to expand to sensitive scenarios such as medical diagnosis, financial risk control, and intelligent transportation, researchers have gradually found that relying solely on distributed training is difficult to completely eliminate risks, because the uploaded gradient information may be reverse-inferred and privacy may be leaked. In this case, how to enhance privacy protection and anti-attack capabilities while maintaining the model's predictive power has become a key issue in promoting the development of distributed intelligent systems. This research is precisely based on this challenge, aiming to build a solution with flexibility and adaptive features by improving the protocol and aggregation mechanism, so as to protect user privacy and resist malicious interference while still maintaining the stability and high-precision prediction of the global model, and promoting its application in sensitive fields such as medical health and financial security.

2. Relevant research

In the field of distributed artificial intelligence model training, when achieving cross-institutional and cross-device collaboration, data privacy protection has always been a core challenge. Due to the fact that participants need to exchange gradients or parameters during the model training process rather than the original data, there is still a potential risk of information leakage. This has prompted researchers to continuously explore how to ensure the security of sensitive information An effective method to maintain model performance as much as possible at the same time. In recent years, federated learning, as a distributed collaboration mechanism that allows each participant to independently train models locally and only share model updates, has gradually become an important technical path to address the security risks brought by centralized data storage. However, the shared gradient information may be reversely analyzed, leading to the leakage of user privacy. Therefore, how to establish a reliable privacy protection system while ensuring training efficiency and model accuracy has become the focus of research attention.

H. Hayati proposed a framework that regards traditional gradient optimization algorithms as dynamical systems and embeds them into high-dimensional target systems[1]. By combining differential privacy methods, it enhances data security while minimizing model performance loss, providing a new theoretical approach for privacy protection in distributed learning environments. In the medical image analysis scenario, the FedDP method designed by L. Pan integrates the differential privacy mechanism into the federated learning framework[2], enabling different medical institutions to complete collaborative training without exchanging the original pathological images. Experiments have proved that this method effectively prevents the leakage of sensitive information that may be caused by gradient reverse inference with only a slight decrease in accuracy. This verified the practical feasibility of differential privacy in protecting medical data. In the application of the industrial field, H. Chen proposed a wind turbine fault prediction method combined with differential privacy. The original features were compressed through the Conv1D-BiLSTM-DAE-TF model, and multiple deep learning methods were adopted for state assessment and anomaly detection[3]. This enables high-precision performance prediction and operation and maintenance optimization of equipment operation data while strictly protecting privacy, fully demonstrating the application value of privacy-preserving federated learning in the management of critical infrastructure.

To address the issue of high computing and communication overhead in cross-institutional and cross-device environments, C. Shen proposed a federated learning framework that combines multiple privacy masks[4], TQRR gradient screening, homomorphic encryption, and secret sharing technologies. This framework significantly reduces computing resource consumption and communication load while ensuring model accuracy. This has enhanced the security and scalability

of the system under large-scale deployment conditions. With the development of 6G networks and large-scale distributed data environments, J. Huang proposed a federated learning method based on conditional generative adversarial networks, which divides the client local network into private modules and public modules, strengthening data isolation and privacy protection, while solving the problems of scarcity and uneven distribution of training data. This has thereby enhanced the training efficiency and generalization ability of the distributed model[5].

These studies have demonstrated diverse strategies for the combination of federated learning and multi-level privacy protection technologies in various application scenarios[6], including differential privacy, system embedding, homomorphic encryption[7], secret sharing, and generative adversarial networks, providing theoretical basis and technical reference for this paper on privacy protection in AI model training and validation in a federated learning environment. At the same time[8], it indicates that by scientifically designing the model structure and privacy mechanism, it is possible to ensure high-performance model training while effectively safeguarding data security[9], providing systematic methods and practical guidance for the construction of future cross-domain privacy-protecting AI systems[10].

3. Research on Modeling and Verification Methods for Privacy Protection in Federated Learning Oriented to Security and Robustness

3.1 Construction of a Secure and scalable Privacy-protecting federated Learning framework

During the collaborative training process of distributed intelligent systems, if there is a lack of effective mechanisms to restrain the behaviors of all parties, it is easy to expose sensitive information due to the gradient upload link, and at the same time, the performance of the global model may also decline due to the malicious tampering of some nodes. To address this dual risk, the framework proposed in the research introduces a multi-role collaboration mechanism at the system level. In this mechanism, an independent institution is responsible for generating and allocating security parameters to ensure the establishment of the initial trust environment. After receiving the key and public parameters, users complete the model update using local data, encrypt and perturb the output results, and then upload them to the central aggregation end. The aggregation end, with the support of cloud-side computing power, completes correlation analysis and robustness optimization, thereby significantly reducing the probability of privacy leakage and poisoning attacks while improving the efficiency of model training. In this collaborative mode, the update parameters of users will undergo consistency measurement and correlation screening before global aggregation. Abnormal gradients are dynamically eliminated or their weights are reduced, while the contributions of trusted nodes are retained and amplified, thereby ensuring that the global model can continuously converge and maintain stable prediction performance.

During the theoretical modeling process, in order to balance privacy protection and model accuracy, the framework introduces a combination of differential processing and order-preserving structured encryption, which enables the uploaded parameters to be masked at the numerical level while still maintaining comparability in order. This design provides feasibility for subsequent anomaly detection. If the local update vector is denoted as g_i, its encrypted and perturbated form can be expressed as formula 1. Here, $\mathrm{Enc}(g_i)$ represents the encryption operation that maintains the ordered structure, and $N(0,\sigma^2)$ is the noise distribution function. By controlling the noise variance σ^2, a balance can be struck between privacy protection and the convergence speed of the model.

$$\tilde{g}_i = \mathrm{Enc}(g_i) + N(0,\sigma^2) \text{(Formula 1)}$$

3.2 Privacy Protection and Robustness Modeling in Federated Learning Environments

The aggregation end relies on powerful computing capabilities to perform consistency measurement, correlation screening, and robustness optimization. While gradually weakening the weight of abnormal gradients, it amplifies the contribution of trusted nodes, enabling the global model to maintain stable convergence characteristics and high prediction performance when dealing with uncertain environments and potential security threats. In the design of the theoretical model, the research chose a combination of differential processing and order-preserving encryption to establish a dynamic balance between privacy protection and model performance.

After receiving the disturbance parameters from each node, the aggregation end will combine the reputation evaluation and correlation test to determine its reliability, and then complete the global integration of the parameters through an adaptive weighting strategy, thereby ensuring that the impact of abnormal information is weakened and the trusted contribution is retained. To verify the performance of this method under different noise levels, multiple comparison scenarios were set up in the experiment to comprehensively compare the accuracy, convergence speed and privacy leakage risk index of the model. The results are shown in Figure 1.

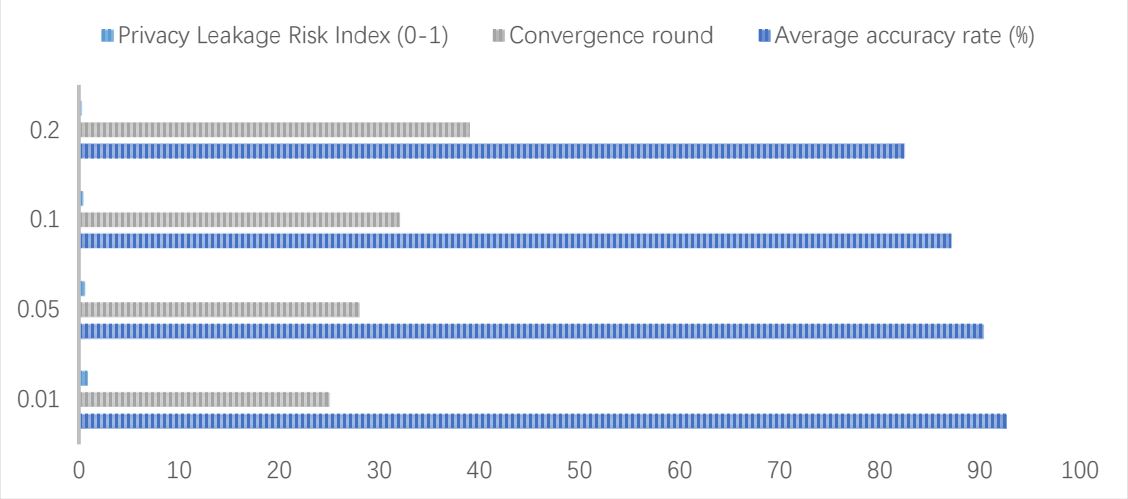

Figure 1 Data performance at different noise levels

It can be clearly seen from Figure 1 that when the noise variance gradually increases, the average accuracy and convergence speed of the model show a significant downward trend, while the privacy leakage risk index continues to decrease. This result fully demonstrates that under this framework, a dynamic trade-off between privacy protection and model performance can be achieved through the adjustment of noise intensity Thus, it still has strong scalability and practicality when facing complex multi-source heterogeneous data and uncertain security threats.

4. Training and validation mechanism of federated learning models for secure collaborative computing

4.1 Model Training and Optimization of Global Aggregation Mechanism

In this system architecture, users act as the main body of distributed computing and independently conduct model training through local private datasets. At the user end, the system generates Gaussian random noise based on the differential privacy parameters set by itself and injects it into the local gradient. Meanwhile, users encrypt the gradient using the mask generated by

the pseudo-random generator. To ensure that local sensitive information is fully protected before being uploaded to the central control node, the entire process effectively combines local data processing with information security. The central control node undertakes the core responsibilities of initializing the global model, generating common parameters, and aggregating gradients. After receiving the encrypted gradients from users, the node decodes and summarizes the gradients based on the preset algorithm, and at the same time adjusts the clipping threshold according to the adaptive mechanism to ensure the convergence and performance of the global model in multiple rounds of training. And it can maintain the balance and reliability of the aggregation results when the gradient contributions of different users vary greatly.

In terms of security threat modeling, the system divides users into honest individuals who follow the protocol and malicious individuals who may carry out attack behaviors. Although the number of malicious users is controlled to be less than half of the total number of users, their potential behaviors may still interfere with the aggregation process of the global model and privacy protection. Therefore, the system design introduces multi-layer protection mechanisms to deal with abnormal situations such as user disconnection, gradient forgery, and mask damage. At the same time, it is assumed that the central control node has the characteristic of "honest but curious", that is, while the node is executing the protocol, it may attempt to infer the user's sensitive information by analyzing the gradient data. This assumption prompts the system to effectively defend against potential internal and external attacks during the data transmission and aggregation process.

In terms of functional design goals, the system takes the protection of user privacy as the premise, while striving to maximize the accuracy of the global model under the premise of reducing the impact of differential privacy noise. During the training process, the user end selects the privacy budget and dynamic clipping threshold that meet personal needs through adaptive algorithms, thereby limiting gradient sensitivity and controlling the amount of noise. To ensure that high-quality gradients are retained as much as possible and fully participate in the update of the global model, the central node aggregates the noisy and encrypted gradients of each user through a secure aggregation mechanism, and uses secret sharing and mask cancellation techniques to restore the updated values of the global model without a trusted third party. This process not only guarantees the accuracy of the training results, It also effectively prevents any single node or attacker from obtaining users' sensitive information through collusion or gradient analysis.

4.2 Privacy Protection and Security Defense under Federated Learning

In a distributed federated learning system, each user node is responsible for independently processing local data and training the model. When performing training tasks, these nodes first calculate gradients using their own private datasets and simultaneously inject random noise into the gradients according to the preset privacy protection policy to prevent sensitive information from being leaked during the upload process. And the gradient is encrypted through the mask generated by the pseudo-random generator, so that the user's data can be fully confidential and protected before the gradient is transmitted to the central aggregation node. During the training process, user nodes can also dynamically adjust the gradient clipping threshold and noise intensity based on their own data characteristics and security requirements, thereby ensuring privacy and security while retaining as much high-value gradient information as possible, so that this information can play a practical role in the iterative update of the global model. As the manager of the global model, the central aggregation node is not only responsible for the initialization of model parameters and the distribution of public information, but also needs to decrypt and weighted summarize the encrypted gradients uploaded by each user node according to the pre-determined algorithm after receiving them. During the calculation process, it dynamically adjusts the aggregation strategy and clipping

threshold based on the differences in gradient contributions of different nodes. To maintain the convergence and stability of the global model in multiple rounds of iterations, while ensuring the model's accuracy and the robustness of training, and to maintain the normal operation of the overall training even when some nodes experience disconnection or abnormal behavior.

To address potential security threats, the system clearly distinguishes between trusted user nodes that follow protocols and malicious nodes that may have abnormal operations during its design, and has established multiple layers of protection measures against potential threats, including verifying gradients, detecting abnormal update behaviors, and ensuring the integrity of masks. At the same time, the system assumes that the central aggregation node has the characteristic of "honesty but curiosity". That is, while strictly implementing the protocol, this node may attempt to infer the user's private data by analyzing the received gradient information. This assumption prompts the entire system to consider the defense strategies against internal and external attacks in every link of gradient transmission, encryption and aggregation, thereby effectively reducing the risk of sensitive information leakage in any case.

In terms of functional implementation, the system aims to maximize the performance of the global model and always prioritizes the protection of user data privacy. User nodes independently select privacy budgets and gradient clipping parameters that suit their own needs through adaptive algorithms, thereby controlling gradient sensitivity and limiting the amplitude of noise injection. This ensures that high-value gradients can fully participate in the update of the global model. After receiving the noisy and encrypted gradients from each user, the central aggregation node securely aggregates them through secret sharing technology and mask cancellation methods, restores the global gradient update values, and completes precise model updates without relying on any trusted third party. The system achieves protection of user privacy, maintenance of model accuracy, defense against collusion behavior, and recovery capability in the event of node disconnection by integrating differential privacy noise injection, gradient mask encryption, and secret sharing mechanisms. This enables distributed federated learning to operate stably and efficiently in a multi-institution, multi-node, and variable network environment. It provides a scientific and reliable technical solution for the training and validation of actual AI models.

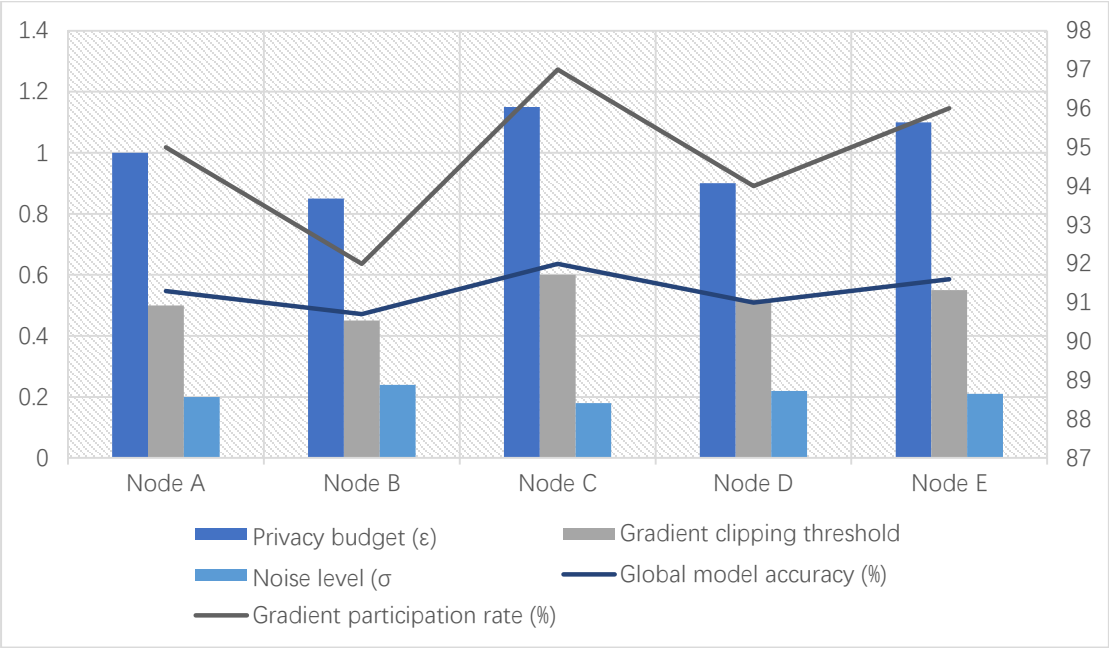

Figure 2 Comparison of model training data

To visually demonstrate the relationship between the privacy protection Settings of user nodes and the model training effect, Figure 2 provides a schematic display of the privacy budget, gradient clipping threshold, noise level of different nodes, as well as the corresponding global model accuracy and gradient participation rate.

It can be observed from the data in Figure 2 that when user nodes reasonably set privacy budgets and gradient clipping parameters, high-value gradients can more fully participate in the update of the global model, thereby maximizing the training accuracy of the model while ensuring user information security, providing a strong reference basis for the actual deployment of multi-node distributed federated learning.

5. Research on Model Training and Validation Methods

Against the backdrop of the rapid development of medical informatization at present, artificial intelligence-assisted diagnosis technology has gradually become a core tool for improving the efficiency and accuracy of diagnosis and treatment. However, due to the high sensitivity of medical data itself, the traditional centralized data processing mode shows significant deficiencies in privacy protection. At the same time, it also limits the ability of different medical institutions to conduct effective data sharing and collaborative training. Therefore, this study proposes a distributed model training method based on federated learning. This method allows each medical institution to independently perform model training while retaining complete patient data, and protects the locally generated model parameters by introducing encryption and random perturbation mechanisms Under the premise of ensuring that sensitive information is not leaked, the processed parameters are safely transmitted to the centralized aggregation node for global model update, thereby achieving multi-institution collaborative learning and knowledge sharing, significantly alleviating the problem of medical data silos, and simultaneously enhancing the predictive ability and application value of the global model.

To verify the feasibility of the proposed method, this study constructed a complete prototype system and applied the theoretical algorithm to actual medical scenarios. The core structure of this system consists of a distributed model training module and an artificial intelligence-assisted diagnosis module. The former is responsible for the local generation, perturbation and encryption of the model in each medical institution. And through a secure aggregation strategy, the training results of various institutions are integrated to generate high-precision global models. In a distributed artificial intelligence-assisted diagnosis system involving multiple institutions, users can obtain complete image analysis and disease prediction services through the encapsulated global model. This process does not require users to directly contact the underlying model or raw data, thereby achieving reliable diagnostic functions while ensuring data privacy and preventing the risk of medical information leakage during transmission and processing. During the model training stage, the system adopts a comprehensive strategy. Each medical institution actively injects random disturbances that conform to statistical characteristics when independently training model parameters locally to encrypt and protect sensitive information. Meanwhile, the aggregation nodes screen, select and correct the uploaded parameters through an optimized aggregation algorithm. This enables the global model constructed in a multi-institution collaborative environment to maintain training stability while taking into account prediction accuracy and possess strong anti-attack capabilities.

The overall system architecture adheres to the dual-system design concept. The front-end subsystem consists of a user interaction interface, a data management module, and a historical record query module. It can provide users with intuitive and operable services such as image upload, diagnosis request submission, and prediction result display. At the same time, it undertakes the

tasks of user access permission control and data management. The terminal system is responsible for centralized management of model training, parameter aggregation, and global model generation. After receiving training requests from various medical institutions, it completes resource scheduling and algorithm execution, and handles abnormal parameters and potential attack behaviors in real time during the model update process to ensure a dynamic balance of efficiency, security, and accuracy of the global model.

Each medical institution can flexibly set training parameters based on its own training objectives, privacy protection level, and dataset characteristics, and send training requests to the back-end system through a secure communication protocol. After parsing the requests, the back-end system triggers local model training and injects random disturbances. Subsequently, the model parameters generated by each institution's training are merged and anomalies are eliminated through a secure aggregation mechanism. Thus, an optimized global model is formed and distributed to various institutions for use. After multiple rounds of iterations, the generalization ability of this global model gradually enhances, while minimizing the potential risk of data leakage.

In the implementation process of artificial intelligence-assisted diagnosis functions, the medical images uploaded by users are first preprocessed by the image processing module, including noise removal and key area segmentation. Then, the system calls the global model to perform disease feature recognition and prediction tasks, and returns the analysis results to the front-end user interface. At the same time, they are synchronously stored in the local database By adopting the model black box encapsulation strategy, users cannot directly access the underlying training data, thus ensuring the interpretability and reliability of the prediction results. The system can effectively deal with abnormal operations or malicious interferences of some nodes in a multi-institution and multi-round training environment. Through local perturbation and secure aggregation mechanisms, it ensures that the global model maintains high-precision prediction capabilities in complex environments, enabling AI-assisted diagnosis to achieve efficient, reliable and secure operations in practical applications.

The method designed in this study has been tested and successfully completed the distributed model training, global model generation and auxiliary diagnosis tasks between various medical institutions and aggregation nodes. When facing potential abnormal behaviors or malicious attacks, the system can adjust the training parameters and aggregation strategies through adaptive algorithms to ensure that the global model achieves the expected goals in terms of robustness and accuracy. Ensure the smooth progress of cross-institutional collaborative training. Overall, the privacy-preserving artificial intelligence model training and validation method based on federated learning proposed in this study not only addresses the limitations of traditional medical AI models in data sharing and privacy protection, but also demonstrates significant application value in enhancing the global model prediction performance, strengthening anti-attack capabilities, and achieving multi-institution collaborative training. It provides a solid technical foundation and feasible implementation plans for the safe deployment of medical artificial intelligence systems in actual environments.

6. Conclusions and Prospects

This paper conducts a systematic study on the protection of users' sensitive data in a non-fully trusted distributed learning environment, and proposes a comprehensive scheme to improve the performance of the global model while ensuring data confidentiality. Through the analysis of the limitations of existing federated learning methods in dealing with malicious gradient injection and privacy leakage, a correlation evaluation mechanism based on gradient rank is designed. Enable the system to effectively distinguish abnormal or malicious gradients even when local differential

privacy noise is added, maintaining the stability and reliability of model updates. To meet the differences among participants in terms of privacy requirements and gradient contributions, a dynamic privacy budget allocation and adaptive gradient clipping strategy are proposed, enabling high-value gradients to have a more significant impact on the global model in each round of training while suppressing the weakening of model performance by noise. In the global model aggregation stage, noise accumulation is reduced through secret sharing and pseudo-random generation techniques, thereby enhancing the accuracy and robustness of distributed training. The feasibility of the method was verified by constructing a distributed system for medical image diagnosis. Experiments show that this method significantly improves the prediction accuracy of the model and the stability of the system while ensuring data privacy, providing a feasible technical solution for achieving intelligent distributed processing and auxiliary decision-making in a highly sensitive data environment. It also lays the foundation for personalized privacy protection and high-performance model training in future multi-task, multi-institutional and cross-domain data scenarios.

References

[1] *Hayati H, Murguia C, Nathan V D W .Immersion and Invariance-based Coding for Privacy-Preserving Federated Learning [J]. 2024.*

[2] *Pan L, Huang M, Wang L, et al. FedDP: Privacy-preserving method based on federated learning for histopathology image segmentation [J]. 2024. DOI:10.1109/BIBM62325. 2024.10822021.*

[3] *Tang X, Wu X, Bao W. Intelligent Prediction-Inventory-Scheduling Closed-Loop Nearshore Supply Chain Decision System [J]. Advances in Management and Intelligent Technologies, 2025, 1(4).*

[4] *Chen H, Hsu H Y, Hsieh J Y, et al. A differential privacy-preserving federated learning scheme with predictive maintenance of wind turbines based on deep learning for feature compression and anomaly detection with state assessment [J].Journal of Mechanical Science & Technology, 2024, 38(7).DOI:10.1007/s12206-024-0616-9.*

[5] *Lu, Z. (2025). AI-Driven Cross-Cloud Operations Language Standardisation and Knowledge Sharing System. European Journal of AI, Computing & Informatics, 1(4), 43-50.*

[6] *Zhang, Xuanrui. "Automobile Finance Credit Fraud Risk Early Warning System based on Louvain Algorithm and XGBoost Model." In 2025 3rd International Conference on Data Science and Information System (ICDSIS), pp. 1-7. IEEE, 2025.*

[7] *Wu, H. (2025). The Commercialization Path of Large Language Models in Start-Ups. European Journal of Business, Economics & Management, 1(3), 38-44.*

[8] *Shen C, Zhang W, Zhou T, et al. An Efficient and Secure Privacy-Preserving Federated Learning Framework Based on Multiplicative Double Privacy Masking [J].Computers, Materials & Continua, 2024, 80(3):4729-4748.DOI:10.32604/cmc.2024.054434.*

[9] *Huang J, Chen Z, Liu S, et al. A Novel Federated Learning Framework Based on Conditional Generative Adversarial Networks for Privacy Preserving in 6G [J].Electronics, 2024, 13(4):20.DOI:10.3390/electronics13040783.*

[10] *Yu, X. (2025). Application Analysis of User Behavior Segmentation in Enhancing Customer Lifetime Value. Journal of Humanities, Arts and Social Science, 9(10).*

[11] *Chen J. Design and Implementation of a Personalized Recommendation System Based on Deep Learning Distributed Collaborative Filtering Algorithm on Social Media Platforms[C]//2025 3rd International Conference on Integrated Circuits and Communication Systems (ICICACS). IEEE, 2025: 1-5.*

[12] Yang D, Liu X. Collaborative Algorithm for User Trust and Data Security Based on Blockchain and Machine Learning [J]. Procedia Computer Science, 2025, 262: 757-765.

[13] Lu, Z. (2025). Design and Practice of AI Intelligent Mentor System for DevOps Education. European Journal of Education Science, 1(3), 25-31.

[14] Zheng, H. (2025). Research on Delay-aware Scheduling Algorithms for Edge Task Migration in High-concurrency Environments. Engineering Advances, 5(4).

[15] Li, J. (2025). The Impact of Distributed Data Query Optimization on Large-Scale Data Processing.