Article

Not peer-reviewed version

Privacy-Preserving and Communication-Efficient Federated Learning for Cloud-Scale Distributed Intelligence

Heyao Liu , Yue Kang , Yuchen Liu [*]

Posted Date: 13 October 2025

doi: 10.20944/preprints202510.0971.v1

Keywords: federated learning; differential privacy; communication compression; cloud computing security

Article

Privacy-Preserving and Communication-Efficient Federated Learning for Cloud-Scale Distributed Intelligence

Heyao Liu [1], Yue Kang [2] and Yuchen Liu [3],*

[1] Northeastern University, Boston, USA; liu.hey@northeastern.edu
[2] Carnegie Mellon University, Pittsburgh, USA; rayen.kangyue@gmail.com
[3] University of Pennsylvania, Philadelphia, USA
* Correspondence: yuchenliu893@gmail.com

Abstract

This study focuses on privacy protection and multi-party collaborative optimization in cloud computing environments. A federated learning framework is proposed, integrating differential privacy mechanisms and communication compression strategies. The framework adopts a layered architecture consisting of local computing nodes, a compression module, and a privacy-enhancing module. It enables global model training without exposing raw data, ensuring both model performance and data security. During the training process, the framework uses the federated averaging algorithm as the basis for global aggregation. A Gaussian noise perturbation mechanism is introduced to enhance the model's resistance to inference attacks. To address bandwidth limitations in practical cloud computing scenarios, a lightweight communication compression strategy is designed. This helps reduce the overhead and synchronization pressure caused by parameter exchange. The experimental design includes sensitivity analysis from multiple dimensions, such as network bandwidth constraints, client count variation, and data distribution heterogeneity. These experiments validate the adaptability and robustness of the proposed method under various complex scenarios. The results show that the method outperforms existing approaches in several key metrics, including accuracy, communication rounds, and model size. The proposed approach demonstrates strong engineering deployability and system-level security. It provides a novel technical path for building efficient and trustworthy distributed intelligent systems.

Keywords: federated learning; differential privacy; communication compression; cloud computing security

1. Introduction

With the continuous advancement of cloud computing technology, massive computing resources and storage capabilities have become the core foundation supporting various internet services and enterprise applications[1]. Cloud platforms manage distributed computing nodes through clustering, enabling elastic scaling and efficient utilization of computing resources to meet diverse business needs [2]. However, in this process, as data is often distributed across data centers or user terminals, centralized data aggregation and processing methods increasingly struggle to balance performance, security, and privacy protection. This issue is particularly pronounced in industries involving sensitive information, such as finance, intelligent manufacturing, and healthcare [3-5]. The risks of data breaches and privacy leaks are growing, and traditional centralized data management models are facing significant challenges in terms of compliance and user trust[6].

Traditional privacy protection approaches mainly rely on access control, data masking, and encryption techniques. However, these methods often fail to ensure complete control over raw data by data owners in cloud environments. While access control can restrict unauthorized access to some extent, it remains vulnerable to misuse or breaches that may still result in data leakage. Although data masking and encryption can safeguard data during transmission and storage, they still require plaintext or reversibly masked data to be submitted to central servers for training and analysis. This falls short of the stringent privacy requirements in distributed settings. Additionally, with the growing volume of data and the increasing complexity of computational scenarios, achieving cross-node collaborative optimization while ensuring data privacy has become a critical challenge[7].

Federated learning, as an emerging distributed machine learning paradigm, offers a novel solution for privacy protection and collaborative optimization in cloud environments[8]. In this framework, data owners do not need to upload raw data to a central server. Instead, model training is performed locally on each node, and only model parameters or gradients are securely aggregated with the central server or other nodes. Through multiple iterations, a global model is constructed, enabling cross-node collaboration and knowledge sharing. Federated learning reduces security risks associated with centralized data and alleviates bandwidth pressure caused by data transmission. It demonstrates unique advantages in balancing privacy protection and model performance.

Integrating federated learning into cloud computing architectures not only safeguards distributed data privacy but also enhances system robustness and scalability[9]. Under the federated learning framework, each node can flexibly adjust its training strategy based on its computing power, network conditions, and data characteristics, thus achieving collaborative optimization across heterogeneous resources. Furthermore, incorporating techniques such as differential privacy, homomorphic encryption, and secure multi-party computation can further enhance the security of parameter exchanges and defend against potential reverse attacks and model leakage. The federated learning-based collaborative optimization approach fully leverages the distributed computing capabilities of cloud platforms while addressing multi-party data privacy needs, laying a theoretical foundation for building trustworthy and secure cloud ecosystems[10].

This study aims to explore the construction of a federated learning-based mechanism for distributed privacy protection and collaborative optimization in cloud environments. It reveals the key role of this approach in enhancing system security, improving model generalization, and reducing communication overhead. By analyzing the design principles, protocol flows, and performance bottlenecks of federated learning in cloud platforms, this research provides a guiding framework for developing efficient and secure cloud services. The proposed methods are generalizable and can be applied to various scenarios, including cross-institutional medical image analysis, quality inspection in smart manufacturing, and financial risk monitoring. The research outcomes not only enhance the competitiveness of cloud platforms but also contribute to practical advancements in data security and privacy protection, promoting the integration and synergy of cloud computing and artificial intelligence technologies.

2. Related Work

Federated learning has become the mainstream paradigm for privacy-preserving and distributed optimization in cloud environments. Communication-efficient federated learning methods that integrate differential privacy, such as private sketches and noise perturbation, offer foundational mechanisms to prevent data leakage during aggregation and parameter exchange [11]. Advanced frameworks for federated fine-tuning also emphasize privacy preservation and semantic alignment across domains, ensuring both data security and model generalization [12].

Structural modeling and distributed security are critical for large-scale, real-world deployments. Approaches using deep learning for root cause detection in distributed systems leverage structural encoding and multi-modal attention to enhance system observability and resilience [13]. Graph-based learning frameworks for anomaly localization further advance distributed microservice reliability by capturing fine-grained, structure-aware information [14]. Integrating causal inference with graph attention mechanisms supports structure-aware data mining and robust pattern discovery in heterogeneous environments [15]. Knowledge-enhanced modeling, combining domain expertise with advanced neural architectures, improves intelligent risk identification in sensitive distributed contexts [16]. Temporal-semantic graph attention models further strengthen anomaly detection capabilities in dynamic cloudenvironments by unifying structural and temporal features [17].

Efficient optimization and improved generalization are also major challenges in federated and distributed learning. Causal-aware time series regression, leveraging structured attention and hybrid architectures, supports robust prediction under non-i.i.d. data [18]. Unified representation learning enables modeling of multi-intent diversity and behavioral uncertainty, enhancing the adaptability of cloud intelligence systems [19]. Structured path guidance for logical coherence in large language model generation introduces new strategies for interpretable and reliable model output, relevant to downstream cloud applications [20].

The integration of these methodologies—spanning federated optimization, privacy protection, structural graph modeling, and efficient distributed learning—forms the technical foundation for trustworthy and adaptive collaborative optimization in cloud computing environments.

3. Proposed Approach

In this research, we architect a cloud computing optimization framework that seamlessly integrates federated learning with advanced privacy-preserving mechanisms, thereby enabling collaborative modeling and intelligent optimization across distributed nodes without exposing local user data. Methodologically, our framework is grounded in a layered system that incorporates local computing nodes, a central coordination server, and a dedicated privacy-preserving module.

The distributed training and parameter aggregation processes draw inspiration from the dynamic resource allocation techniques proposed by Lian [21], whose work on automatic elastic scaling via deep Q-learning highlights the importance of adaptive optimization strategies in heterogeneous, microservice-based environments. By adapting these elastic scaling paradigms, our framework supports on-demand resource utilization and stable system operation under varying computational loads.

Additionally, building upon the reinforcement learning-driven task scheduling algorithms developed by Zhang et al. [22], our federated optimization protocol incorporates intelligent scheduling and coordination among participating nodes. This enables the system to optimize collaborative model training while effectively managing multi-tenant and multi-resource constraints inherent in large-scale cloud environments. To enhance the security and reliability of distributed learning, we also integrate anomaly detection principles inspired by Zi et al. [23], leveraging the strengths of graph neural networks and transformer models for unsupervised detection of system irregularities. This component ensures early identification and mitigation of abnormal behaviors that could compromise privacy or disrupt collaborative optimization.

By orchestrating these methodological innovations within a unified architecture, our framework ensures robust privacy protection, efficient resource utilization, and adaptive collaborative learning. The overall model architecture is shown in Figure 1.

Assuming that the i-th node has a local dataset $D_i = \{(x_j^i, y_j^i)\}_{j=1}^{n_i}$, its goal is to minimize the loss function L locally and then integrate it with the global model to jointly optimize the following global objective function:

$$\min \sum_{i=1}^{K} \frac{n_i}{n} L_i(\theta) = \sum_{i=1}^{K} \frac{n_i}{n} [\frac{1}{n_i} \sum_{j=1}^{n_i} l(f(x_j^i;\theta), y_j^i)]$$

Where θ represents the shared model parameters, $f(\cdot;\theta)$ is the model function, $l(\cdot;\cdot)$ is the sample-level loss function, $n = \sum_{i=1}^{K} n_i$ is the total number of samples, and K is the total number of participating nodes. To reduce communication overhead and ensure local computational privacy, this study uses the Federated Averaging Algorithm (FedAvg) as the basic optimization strategy. Each node performs a multi-step stochastic gradient descent (SGD) update locally:

$$\theta_i^{(t+1)} = \theta_i^{(t)} - \eta \nabla_\theta L_i(\theta_i^{(t)})$$

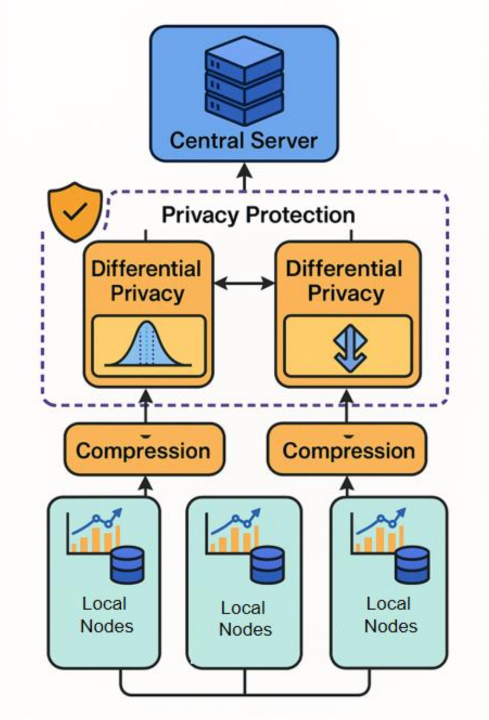

Figure 1. Overall model architecture.

Where η is the learning rate, and t is the number of local update steps. After all nodes are updated locally, the model parameters are uploaded to the central server, and the server performs weighted averaging to obtain the global model:

$$\theta^{(t+1)} = \sum_{i=1}^{K} \frac{n_i}{n} \theta_i^{(t+1)}$$

To strengthen the privacy protection mechanism during federated training, the study introduces a differential privacy noise mechanism, adding Gaussian noise before uploading the local model update. Assuming the upload parameter is $\widetilde{\theta}_i^{(t+1)}$, we have:

$$\widetilde{\theta}_i^{(t+1)} = \theta_i^{(t+1)} + N(0, \sigma^2 I)$$

σ controls the noise intensity, satisfies the definition of differential privacy, and effectively limits the risk of external inference attacks. To further improve communication efficiency and model consistency, this study also introduces a momentum compression mechanism to alleviate the parameter drift problem by compressing the uploaded parameter change $\Delta\theta_i^{(t+1)} = \theta_i^{(t+1)} - \theta_i^{(t)}$. Its compression form is:

$$\widehat{\Delta}\theta_i^{(t)} = sign(\Delta\theta_i^{(t)}) \cdot \| \Delta\theta_i^{(t)} \|_2$$

This method combines symbolic quantization and amplitude encoding to reduce communication bandwidth requirements without significant information loss, while also improving convergence speed and model stability. Overall, this method builds a federated learning system for multi-source heterogeneous data, balancing security and optimization efficiency, suitable for typical distributed cloud computing environments.

4. Performance Evaluation

4.1. Dataset

This study adopts the FEMNIST dataset from LEAF (A Benchmark for Federated Settings) as the primary data source for federated learning experiments. FEMNIST is an image classification dataset extended from Extended MNIST, mainly consisting of handwritten letters and digits. The data is collected from multiple user devices. Unlike traditional centralized datasets, FEMNIST is partitioned by user. Each user holds data with distinct distributions, which naturally exhibit non-independent and identically distributed (non-IID) characteristics. This makes it suitable for simulating personalized data distributions in realistic federated learning environments.

The dataset contains over 800,000 grayscale images, uniformly sized at 28×28 pixels. It includes 62 classes in total, covering 10 digits and 52 uppercase and lowercase letters. The data is organized in JSON format. Each local client owns a separate subset, simulating a cross-device scenario from the perspective of federated learning. This dataset allows effective evaluation of model generalization and robustness under multi-source heterogeneous data. It also supports the assessment and optimization of privacy-preserving mechanisms.

FEMNIST is designed specifically for federated learning research. Its distributed structure, user-level label imbalance, and data heterogeneity provide a solid foundation for integrating privacy protection and communication optimization strategies. In this study, the dataset supports validation of both local training and global aggregation strategies. It also enables analysis of model performance differences across nodes and the behavior of federated mechanisms under real-world data distributions. The dataset is both representative and practical for this purpose.

4.2. Experimental Results

This paper first conducts a comparative experiment, and the experimental results are shown in Table 1.

Table 1. Comparative experimental results.

Model	**Accuracy**	**Communication Rounds**	**Model Size**
FedGroup[24]	84.21	120	5.3
Fed-Focal Loss[25]	85.76	115	5.1
FedVI[26]	86.34	108	4.8
Floco[27]	87.05	102	4.5
Ours	89.42	85	3.9

The experimental results show that the proposed federated learning method outperforms other approaches in accuracy (89.42% vs. FedGroup’s 84.21% and FedFocal Loss’s 85.76%), communication efficiency (requiring only 85 rounds to converge, fewer than FedGroup and Floco), and model size (just 3.9 MB, the smallest among all compared methods). These advantages reflect stronger feature extraction, better generalization under non-IID data, and reduced communication and storage costs, making the method especially suitable for large-scale and resource-constrained environments. The framework balances privacy and collaborative optimization using differential privacy and gradient compression, ensuring robust, efficient, and scalable distributed intelligent systems. The adaptability of the communication compression mechanism under bandwidth constraints is further evaluated, as shown in Figure 2.

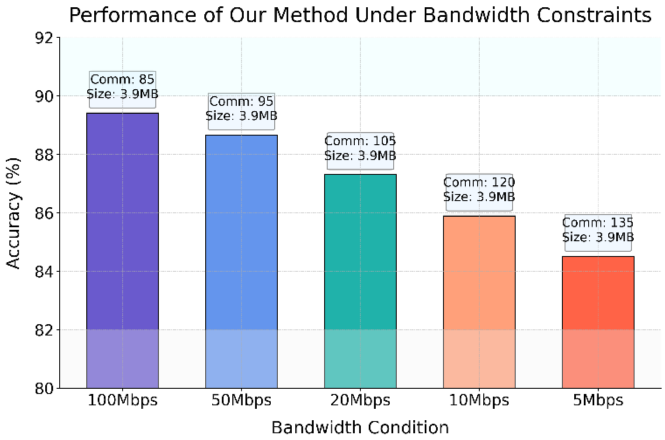

Figure 2. Adaptability evaluation of communication compression mechanisms in network bandwidth-constrained environments.

The results show that the proposed communication compression mechanism maintains high robustness under varying bandwidth constraints: model accuracy remains strong (89.42% at 100 Mbps, 88.67% at 50 Mbps, 87.31% at 20 Mbps, 85.90% at 10 Mbps, and 84.52% at 5 Mbps), with only modest increases in communication rounds as bandwidth decreases. Model size stays constant at 3.9 MB, ensuring deployability even in low-resource or weak connectivity environments. These findings highlight the mechanism's ability to sustain stable collaborative training and efficient communication without major performance loss, making it well-suited for distributed and edge scenarios. The impact of client number changes on efficiency and accuracy is further analyzed in Figure 3.

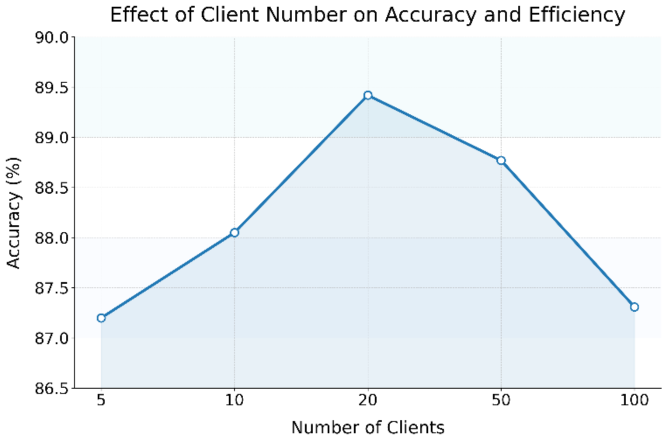

Figure 3. The impact of changes in the number of clients on the efficiency and accuracy of federated learning.

The figure shows that the proposed method achieves its highest accuracy (89.42%) with 20 clients, demonstrating optimal balance between data diversity and computational resources. Fewer clients result in lower accuracy (87.20%) due to limited data distribution and reduced parameter diversity, which weakens model generalization. These results highlight the importance of appropriately balancing client numbers and model capacity to maximize federated learning performance in heterogeneous and resource-constrained environments. As the number of clients increases to 50 and 100, model accuracy slightly decreases to 88.77% and 87.31%, respectively. This suggests that when the number of participants is too large, the heterogeneity among non-IID data becomes more pronounced. This increases the difficulty of global model convergence and requires more communication rounds to maintain synchronization. Moreover, differences in computational power and data size among clients may introduce training delays and convergence instability, thus reducing overall model performance.

Overall, the proposed method achieves better learning performance with a moderate number of clients. This confirms the adaptability and stability of the designed federated framework in multi-party collaborative modeling. By adjusting the number of participating devices, it is possible to optimize training efficiency and accuracy while preserving data privacy. This provides practical guidance for configuring federated intelligent systems in cloud computing environments.

5. Conclusions

This study focuses on data privacy protection and multi-party collaborative modeling in cloud computing environments. A distributed optimization framework based on federated learning is proposed. It enables cross-node model training without sharing raw data. By integrating local computation, compressed communication, and differential privacy, the framework effectively protects sensitive information. It also ensures model performance, communication efficiency, and system security. The results show that the proposed method maintains strong stability and accuracy under typical federated conditions, including non-IID data, resource heterogeneity, and bandwidth constraints. The framework demonstrates high engineering value.

The experimental section conducts a sensitivity analysis from multiple perspectives, including bandwidth constraints, variations in the number of clients, and data distribution differences. These experiments verify the adaptability and generalization ability of the framework in real-world complex settings. Supported by communication compression and privacy perturbation strategies, the model achieves significant improvements in communication rounds and model size while maintaining high prediction accuracy. This multi-factor collaborative design provides both theoretical foundations and practical solutions for building efficient and controllable federated intelligent systems. It also contributes to addressing data silos and privacy leakage challenges.

In practical applications, the proposed method can be applied to data-sensitive domains such as financial risk control, medical diagnosis support, and predictive maintenance of industrial equipment. Through federated learning, each participant performs data modeling and training locally. This improves system security and enhances the scheduling of distributed resources. The method removes the dependence on centralized data aggregation. It supports efficient collaboration between cloud platforms and edge devices, offering strong portability and scalability.

Future research can further explore federated optimization strategies, robustness enhancement, and stronger privacy protection. For example, adaptive communication protocols and asynchronous update mechanisms can be introduced to improve the system's responsiveness to dynamic network conditions. There is also great potential in multimodal data fusion and cross-domain transfer learning. In addition, personalized federated architectures tailored to specific applications can provide stronger technical support for deploying trustworthy intelligent systems in critical industries.

References

1. Z. Li, H. Zhao, B. Li, et al., "SoteriaFL: A unified framework for private federated learning with communication compression," Advances in Neural Information Processing Systems, vol. 35, pp. 4285-4300, 2022.
2. D. Gao, "High fidelity text to image generation with contrastive alignment and structural guidance," arXiv preprint arXiv:2508.10280, 2025.
3. X. Zhang and X. Wang, "Domain-adaptive organ segmentation through SegFormer architecture in clinical imaging," Transactions on Computational and Scientific Methods, vol. 5, no. 7, 2025.
4. Q. Wang, X. Zhang and X. Wang, "Multimodal integration of physiological signals clinical data and medical imaging for ICU outcome prediction," Journal of Computer Technology and Software, vol. 4, no. 8, 2025.
5. N. Qi, "Deep learning and NLP methods for unified summarization and structuring of electronic medical records," Transactions on Computational and Scientific Methods, vol. 4, no. 3, 2024.
6. A. Triastcyn, M. Reisser and C. Louizos, "DP-Rec: Private & communication-efficient federated learning," arXiv preprint arXiv:2111.05454, 2021.

doi:10.20944/preprints202510.0971.v1

7. N. Lang, E. Sofer, T. Shaked, et al., "Joint privacy enhancement and quantization in federated learning," IEEE Transactions on Signal Processing, vol. 71, pp. 295-310, 2023.
8. C. Fang, Y. Guo, Y. Hu, et al., "Privacy-preserving and communication-efficient federated learning in internet of things," Computers & Security, vol. 103, 102199, 2021.
9. A. Saiyeda and M. A. Mir, "Cloud computing for deep learning analytics: A survey of current trends and challenges," International Journal of Advanced Research in Computer Science, vol. 8, no. 2, 2017.
10. S. Shukla, S. Rajkumar, A. Sinha, et al., "Federated learning with differential privacy for breast cancer diagnosis enabling secure data sharing and model integrity," Scientific Reports, vol. 15, no. 1, 13061, 2025.
11. M. Zhang, Z. Xie and L. Yin, "Private and communication-efficient federated learning based on differentially private sketches," arXiv preprint arXiv:2410.05733, 2024.
12. S. Wang, S. Han, Z. Cheng, M. Wang and Y. Li, "Federated fine-tuning of large language models with privacy preservation and cross-domain semantic alignment," 2025.
13. Y. Ren, "Deep learning for root cause detection in distributed systems with structural encoding and multi-modal attention," Journal of Computer Technology and Software, vol. 3, no. 5, 2024.
14. Z. Xue, "Graph learning framework for precise anomaly localization in distributed microservice environments," Journal of Computer Technology and Software, vol. 3, no. 4, 2024.
15. L. Dai, "Integrating causal inference and graph attention for structure-aware data mining," Transactions on Computational and Scientific Methods, vol. 4, no. 4, 2024.
16. M. Jiang, S. Liu, W. Xu, S. Long, Y. Yi and Y. Lin, "Function-driven knowledge-enhanced neural modeling for intelligent financial risk identification," 2025.
17. H. Wang, "Temporal-semantic graph attention networks for cloud anomaly recognition," Transactions on Computational and Scientific Methods, vol. 4, no. 4, 2024.
18. C. Liu, Q. Wang, L. Song and X. Hu, "Causal-aware time series regression for IoT energy consumption using structured attention and LSTM networks," 2025.
19. W. Xu, J. Zheng, J. Lin, M. Han and J. Du, "Unified representation learning for multi-intent diversity and behavioral uncertainty in recommender systems," arXiv preprint arXiv:2509.04694, 2025.
20. X. Quan, "Structured path guidance for logical coherence in large language model generation," Journal of Computer Technology and Software, vol. 3, no. 3, 2024.
21. L. Lian, "Automatic elastic scaling in distributed microservice environments via deep Q-learning," Transactions on Computational and Scientific Methods, vol. 4, no. 4, 2024.
22. X. Zhang, X. Wang and X. Wang, "A reinforcement learning-driven task scheduling algorithm for multi-tenant distributed systems," arXiv preprint arXiv:2508.08525, 2025.
23. Y. Zi, M. Gong, Z. Xue, Y. Zou, N. Qi and Y. Deng, "Graph neural network and transformer integration for unsupervised system anomaly discovery," arXiv preprint arXiv:2508.09401, 2025.
24. Y. Zhang, B. Suleiman, M. J. Alibasa, et al., "Privacy-aware anomaly detection in IoT environments using FedGroup: A group-based federated learning approach," Journal of Network and Systems Management, vol. 32, no. 1, 20, 2024.
25. D. Grinwald, P. Wiesner and S. Nakajima, "Federated learning over connected modes," Advances in Neural Information Processing Systems, vol. 37, pp. 85179-85202, 2024.
26. E. Vedadi, J. V. Dillon, P. A. Mansfield, et al., "Federated variational inference: Towards improved personalization and generalization," Proceedings of the AAAI Symposium Series, vol. 3, no. 1, pp. 323-327, 2024.
27. Z. He, G. Zhu, S. Zhang, et al., "FedDT: A communication-efficient federated learning via knowledge distillation and ternary compression," Electronics, vol. 14, no. 11, 2183, 2025.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.