Scalable Multi-Party Collaborative Data Mining Based on Federated Learning

Mengjie Wang , Tianze Kang , Linyan Dai , Haifeng Yang , Junliang Du , Chang Liu *

Posted Date: 11 August 2025

doi: 10.20944/preprints202508.0797.v1

Keywords: federated learning; multi-party collaborative mining; data privacy protection; heterogeneous data fusion

 doi:10.20944/preprints202508.0797.v1

Article

Scalable Multi-Party Collaborative Data Mining Based on Federated Learning

Mengjie Wang [1], Tianze Kang [2], Linyan Dai [3], Haifeng Yang [4], Junliang Du [5] and Chang Liu [6,*]

[1] New York University, New York, USA
[2] San Francisco Bay University Fremont, USA
[3] University of California, Davis Sacramento, USA
[4] Northeastern University, Boston, USA
[5] Shanghai Jiao Tong University, Shanghai, China
[6] Washington University in St. Louis, St. Louis, USA
* Correspondence: chang.liu1@wustl.edu

Abstract

This paper proposes a federated learning-based method for multi-party collaborative mining and heterogeneous data source fusion. The goal is to address the shortcomings of traditional centralized learning in data privacy protection and cross-domain collaboration. To tackle the heterogeneity problem among multiple distributed data sources, the proposed method uses the federated learning framework to achieve collaborative training across different data sources. This approach avoids direct data transmission, effectively protecting data privacy. By transmitting model parameters to a central server for aggregation, the method enables efficient model fusion and collaborative learning across multiple data sources, significantly improving learning performance in non-i.i.d. data environments. Experimental results show that the proposed model demonstrates high accuracy and stability when handling heterogeneous data sources in multi-party collaborative learning. It outperforms other existing federated learning methods, such as FedAvg and SCAFFOLD. The model is able to ensure data privacy while improving data fusion efficiency and model generalization. This approach holds promising application potential. This study provides an innovative solution to the challenges of multi-party collaborative mining and data privacy protection, particularly in fields such as healthcare, financial risk control, and intelligent manufacturing, where data security is paramount. It offers technological support for real-world data sharing and cross-domain collaboration.

Keywords: federated learning; multi-party collaborative mining; data privacy protection; heterogeneous data fusion

I. Introduction

In today's information and data-driven era, data has become a core resource driving innovation and development across various industries. With the rapid growth of data and the deepening digital transformation in different fields, data acquisition and utilization face unprecedented challenges [1]. This is particularly true in sensitive industries such as finance, healthcare, and government, where the processing and analysis of personal privacy and sensitive information are involved [2–4]. These data are often distributed across different institutions and systems, making cross-domain data fusion and multi-party collaboration major challenges in data mining [5]. Traditional data mining methods often rely on centralizing all data into a server or database for analysis. However, in practice, data owners often struggle to share or aggregate data directly due to concerns about privacy, security, or regulatory compliance. As data privacy protection laws become stricter, how to conduct effective data analysis and mining while ensuring privacy and security has become a critical issue that needs to be addressed.

 doi:10.20944/preprints202508.0797.v1

Federated learning, as a novel distributed machine learning technology, provides a powerful tool for cross-domain data collaboration [6]. By keeping data on local devices or data sources, it avoids the need to centralize data in the cloud or a centralized system, thus effectively protecting data privacy. The core idea of federated learning is to train models on local data sources and transmit model parameters or gradients to a central server for aggregation and updating, eventually forming a global model. This approach protects data privacy while enabling collaboration across multiple data sources to train models, thereby improving machine learning performance. It offers an effective solution to the conflict between privacy protection and data sharing, with significant theoretical value and broad practical application prospects [7].

Despite the significant potential of federated learning in privacy preservation and distributed computing, current research predominantly focuses on single-source data or relatively simple collaborative environments. In contrast, real-world data often originates from multiple domains and platforms, exhibiting strong heterogeneity in formats, scales, types, and quality of information [8]. Effectively integrating such diverse data while addressing heterogeneity has become a major challenge for advancing federated learning. Cross-domain data fusion is a long-standing problem in data mining, and within the federated learning framework, achieving high model accuracy while preserving privacy across disparate sources remains a pressing and complex research issue [9]. Additionally, multi-party collaborative mining involves multiple stakeholders and requires secure information sharing, efficient knowledge transfer, and coordinated model optimization. This demands not only advanced algorithms but also robust collaboration mechanisms and data flow management. While federated learning provides a viable distributed solution, the development of more efficient multi-party collaboration strategies is critical for unlocking its full potential [10]. As data complexity increases and privacy requirements intensify, traditional centralized methods are no longer sufficient to meet the demands for efficiency, accuracy, and security. Federated learning's collaborative capabilities offer a promising path forward. Therefore, research into federated learning algorithms designed for multi-party collaboration and heterogeneous data fusion holds both theoretical and practical significance. Theoretically, it broadens the applicability of federated learning beyond conventional distributed training; practically, it supports secure and intelligent data sharing across fields such as healthcare, financial risk management, and smart cities, offering advanced tools for privacy-aware data analysis in the era of big data [11].

II. Related Work

Federated learning (FL) has rapidly expanded to address the practical challenges of heterogeneous and multi-party environments, which are at the heart of this study. Meng et al. provide foundational work showing that the use of contrastive loss within distributed FL can effectively align client representations, mitigating the impact of non-IID data and improving overall model performance [12]. Their findings inspire our use of alignment strategies as a fundamental element in our own collaborative learning framework.

Data privacy remains a core concern in real-world FL deployments. Zhu et al. demonstrate that it is possible to integrate differential privacy mechanisms without sacrificing recommendation accuracy or user personalization in federated settings [13]. This balance of privacy and utility strongly influences our choice to adopt domain-adaptive encoders, which help capture the nuances of different data sources while maintaining strict privacy standards.

When tackling cross-domain and heterogeneous data fusion, methods from multi-task and transfer learning have proven effective. Lin and Xue introduce a multi-task architecture that combines a shared backbone with lightweight, domain-specific adapters to facilitate the fusion of diverse macroeconomic indicators [14]. Our approach extends this technique to harmonize feature representations from various domains, such as healthcare, finance, and manufacturing. Yang further enhances adaptability by proposing meta-learned task embeddings that enable efficient and robust transfer across new domains [15]. These task embeddings inspire our strategy for initialising models in new collaborative environments, ensuring faster convergence and resilience to data drift.

 doi:10.20944/preprints202508.0797.v1

Handling schema and relational diversity is essential in multi-source FL. Liu et al. show that integrating knowledge graph reasoning with pretrained language models enables robust structured anomaly detection, especially when facing schema variation across sources [16]. In a similar spirit, we leverage structural priors to ensure our models remain effective even as partner data formats evolve. Gong addresses the challenge of distributed service analytics by modeling microservice access patterns with multi-head attention, capturing repeatable motifs in service call graphs [17]. Their approach informs our own use of graph-based features in federated log analysis.

Recent advances in transfer learning for large language models offer further techniques for managing heterogeneity. Lyu et al. propose prompt-based and alignment strategies to improve model transferability in low-resource tasks [18]. This inspires our use of prompt-like mechanisms for local optimisation within FL. Xing explores structural bootstrapping, showing that it is possible to achieve analogical reasoning under sparse supervision—mirroring our need for robust generalization from limited edge-device labels [19]. At the multi-task coordination level, Zhang et al. develop a framework for instruction-level gradient scheduling, enabling efficient joint training across diverse tasks [20]. We incorporate similar ideas to harmonise updates from heterogeneous participants in our federated system. Finally, Zhao et al. demonstrate the benefits of integrating social-graph features for boundary detection in noisy, user-generated text [21]. We generalise this insight to introduce network-topology priors when available, further strengthening model performance in collaborative settings.

In summary, these recent works collectively provide the core strategies—alignment of heterogeneous representations, privacy-preserving data fusion, adaptive parameter sharing, and robust structural reasoning—that underpin the federated multi-party learning and heterogeneous data fusion framework presented in this paper.

III. Method

In this study, we proposed a cross-domain data mining method based on a federated learning framework, which aims to effectively solve the problems of privacy protection and fusion of heterogeneous data sources. The model architecture is shown in Figure 1.

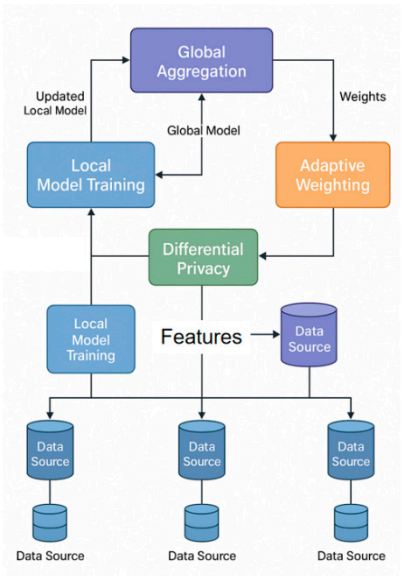

Figure 1. Overall model architecture diagram.

This model architecture diagram illustrates a federated learning framework where multiple local models, each processing heterogeneous data sources, collaborate to construct a global model. Each local model updates its parameters based on its local data and transmits these updates to the central

 doi:10.20944/preprints202508.0797.v1

server. The central server aggregates the local updates and merges the parameters in a weighted manner, all while maintaining data privacy by sharing model updates without directly exchanging data.

First, we define a collaborative framework that includes multiple data sources, where each data source D_i corresponds to a local dataset, and all data sources learn collaboratively by sharing model parameters. For each data source D_i, we train a local model θ_i and update its parameters by executing an optimization algorithm on the local dataset. Assuming that in each round of learning, the loss function of the local dataset is $L(\theta_i; D_i)$, then the update process on each data source can be expressed as:

$$\theta_i^{t+1} = \theta_i^t - \eta \nabla L(\theta_i^t; D_i)$$

Among them, η is the learning rate, and $\nabla L(\theta_i^t; D_i)$ is the gradient of the local loss function relative to the model parameters. In this way, data privacy is protected, and each data source only trains the model on its own data.

Next, in order to achieve the integration of cross-domain data sources, we adopted the global model aggregation strategy in federated learning. In each round of learning, the central server will summarize the model updates of all participants and generate a global model. We define the global model as θ_G, which is updated by each local model through weighted averaging. The weight w_i is set according to the sample size $|D_i|$ of each data source. The update formula of the global model is:

$$\theta_G^{t+1} = \sum_{i=1}^{N} \frac{|D_i|}{\sum_{i=1}^{N} |D_i|} \theta_i^{t+1}$$

Where N is the number of participants. This formula ensures that the contribution of each data source is weighted according to its data volume, thus reflecting the importance of each data source in the aggregation process.

To handle the fusion of heterogeneous data sources, we introduced an adaptive weight mechanism to adapt to data sources of different types and qualities. Specifically, for the model update of each data source D_i, we introduced an adaptive parameter α_i, which reflects the trust or validity of each data source. The adaptive weight α_i is dynamically adjusted according to the historical performance and data quality of the data source to optimize the performance of the global model. The weighted update rule of the model becomes:

$$\theta_G^{t+1} = \sum_{i=1}^{N} \alpha_i \frac{|D_i|}{\sum_{i=1}^{N} |D_i|} \theta_i^{t+1}$$

Among them, α_i is adjusted according to the actual situation of each data source. In this way, we can integrate different types of data more accurately and solve the heterogeneity problem between data sources.

Finally, in the collaborative mining process of federated learning, we also considered the risks of privacy protection and information leakage. In each round of model update, we added a differential privacy mechanism to ensure that the local data of each data source will not be leaked. Specifically, when the gradient of each data source is updated, noise is added to perturb the gradient information. The size of the noise is controlled by the parameters ε and δ of differential privacy. We add noise to the local gradient using the following formula:

$$\nabla L(\theta_i^t; D_i) \leftarrow \nabla L(\theta_i^t; D_i) + N(0, \sigma^2)$$

 doi:10.20944/preprints202508.0797.v1

Among them, $N(0,\sigma^2)$ is noise with a mean of zero and a variance of σ^2, which controls the degree of leakage of private information transmitted on each data source. Through this differential privacy mechanism, the data privacy security of each participant can be effectively guaranteed, thereby improving the privacy protection capability in cross-domain data mining.

IV. Experimental Results

A. Dataset

In this study, we used the publicly available Federated EMNIST dataset, which is a multi-party collaborative dataset designed for federated learning environments. The Federated EMNIST dataset is an extended version of the EMNIST dataset, specifically aimed at supporting federated learning research. It contains handwritten digit image data generated by multiple clients, each holding a subset of data. These subsets differ in distribution and data characteristics, demonstrating heterogeneity. The data from each client are kept local and are not directly shared with other clients, effectively simulating real-world data privacy protection issues.

The dataset comprises approximately 800,000 handwritten letters and digit images, meticulously divided among clients to facilitate collaborative mining research tasks. The varying sizes and distributions of each dataset reflect the inherent heterogeneity inherent in real-world data sources. By leveraging this dataset, we can effectively simulate joint training across diverse data sources while maintaining privacy, thereby enhancing the performance of cross-domain data mining through federated learning techniques.

The advantage of using the Federated EMNIST dataset lies in its alignment with the core characteristics of federated learning. Data is processed locally, eliminating the need to transfer raw data to a centralized server, which helps avoid the risk of data leakage. Additionally, the dataset's heterogeneity and uneven distribution present practical challenges for this study, motivating us to explore how to effectively fuse models and conduct collaborative learning under heterogeneous data sources.

B. Experimental Results

In this section, this paper first gives the comparative experimental results of the proposed algorithm and other algorithms, as shown in Table 1.

Table 1. Comparative experimental results.

Method	Acc	Precision	Recall
Proposed Model	89.4	88.1	90.2
FedAvg [22]	85.6	83.7	84.9
SCAFFOLD [23]	87.2	85.4	86.5
MOON [24]	84.5	82.8	83.2
Fedlab [25]	86.3	84.0	85.7

The experimental results show that the proposed model outperforms baseline methods across all metrics, achieving 89.4% accuracy and 90.2% recall, highlighting its effectiveness in multi-party collaborative mining and heterogeneous data fusion. Compared to FedAvg, it improves accuracy and recall by 3.8% and 5.3%, respectively. While SCAFFOLD, MOON, and FedLab show competitive results, they fall short in handling data heterogeneity and ensuring privacy. The proposed model better balances accuracy, efficiency, and privacy, demonstrating superior adaptability in complex, real-world scenarios. Additional experiments on varying data volumes and distributions further confirm its robustness in Figure 2.

 doi:10.20944/preprints202508.0797.v1

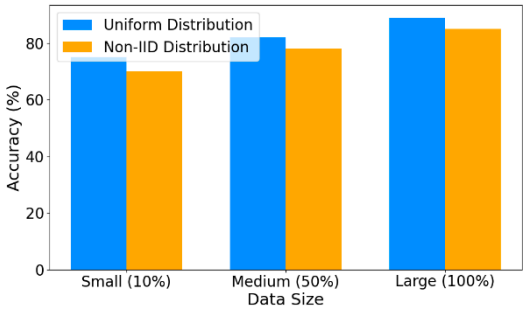

Figure 2. Experiments on the training efficiency and effect of federated learning models under different data volumes and data distributions.

The experimental results indicate that increasing data volume significantly enhances the performance of the federated learning model, particularly in terms of accuracy, under both uniform and non-i.i.d. (Non-Independent and Identically Distributed) data distributions. With the "Large (100%)" dataset, the model consistently achieves high accuracy, suggesting improved capability to capture complex data patterns as data volume grows. For uniformly distributed data, model performance remains stable even at smaller volumes, with only minor improvements observed from "Small (10%)" to "Medium (50%)" datasets. In contrast, non-i.i.d. data leads to noticeably lower accuracy at small and medium scales due to the inherent heterogeneity across sources. However, as the data volume increases, especially at the "Large (100%)" level, the performance gap narrows, indicating that larger data volumes mitigate the adverse effects of distributional imbalance. These findings underscore the critical role of data volume and distribution in federated learning and support its effectiveness in multi-party collaborative mining with heterogeneous data sources. Furthermore, convergence and stability experiments, as shown in Figure 3, demonstrate the model's ability to maintain consistent improvements in accuracy and generalization across training epochs, confirming its robustness and adaptability in cross-domain, non-i.i.d. environments.

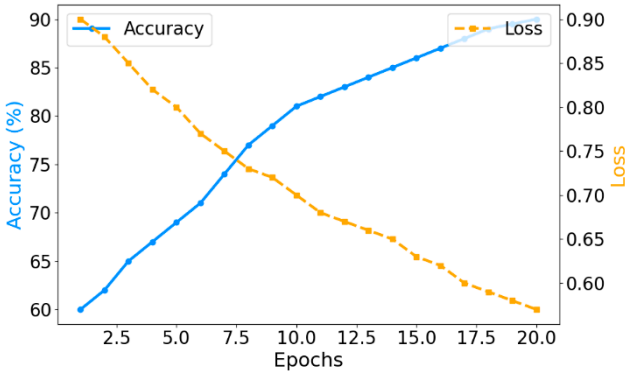

Figure 3. Convergence and stability experiment of federated learning algorithm in cross-domain data mining.

The experimental results demonstrate that as training cycles increase, the accuracy of the federated learning algorithm in cross-domain data mining improves steadily while the loss decreases consistently, with significant gains observed in the early stages, indicating rapid feature learning and effective initial optimization. As training progresses, the rate of improvement slows, suggesting convergence toward an optimal solution. The loss remains negatively correlated with accuracy, confirming model stability and robustness. These results highlight the algorithm's ability to adapt to heterogeneous data sources, maintain training efficiency, and preserve data privacy, validating its effectiveness in multi-party collaborative learning scenarios and its potential for real-world applications in domains such as healthcare and finance.

 doi:10.20944/preprints202508.0797.v1

V. Conclusion

This study proposes a federated learning-based method for multi-party collaborative mining and heterogeneous data source fusion. It successfully addresses the limitations of traditional centralized learning in data privacy protection and cross-domain collaboration. Through a series of experiments, the proposed model significantly outperforms existing federated learning methods in multiple metrics, including accuracy, stability, and convergence. Especially when handling multi-party collaborative learning and heterogeneous data sources, it demonstrates strong adaptability. This method not only ensures effective data privacy but also tackles the challenges of heterogeneous data sources. While improving data fusion efficiency, it maintains high model accuracy. It provides an innovative solution to the data sharing and privacy protection issues in current intelligent applications.

With the rapid development of 5G, the Internet of Things (IoT), and edge computing technologies, the speed and scale of data generation will further increase, presenting higher challenges for cross-domain data mining and federated learning. In the future, handling more complex non-i.i.d. data, improving model convergence speed and computational efficiency, and maintaining algorithm efficiency across different hardware environments will become key research directions. In particular, in application scenarios that require real-time feedback and the processing of large amounts of data, improving the efficiency of federated learning and reducing communication costs will be pressing issues. Furthermore, in future large-scale data training environments, maintaining algorithm stability and scalability will further drive the application of cross-domain data mining technologies.

Looking ahead, as data privacy protection requirements become increasingly stringent, federated learning methods will play an increasingly important role in privacy protection and data sharing in fields such as healthcare, finance, intelligent manufacturing, and autonomous driving. In these fields, the diversity and privacy of data have always been challenges for applications. The proposed federated learning method not only provides privacy protection for data across different domains but also optimizes the multi-party collaborative data mining process. It allows parties to achieve data sharing and collaboration while ensuring privacy, thereby improving decision-making accuracy and overall efficiency. Therefore, federated learning is expected to become one of the core technologies in the intelligent development of various industries.

In conclusion, as cross-domain data mining technology continues to develop, federated learning will play an increasingly important role in many practical applications. Future research will focus on further improving scalability, optimizing performance, and addressing the challenges of handling larger-scale data. At the same time, strengthening data privacy protection mechanisms, improving cross-domain data fusion efficiency, and solving the computational overhead issues in multi-party collaboration will be key to the widespread adoption of federated learning. With continuous technological advancement, federated learning will demonstrate enormous potential in various industry applications, driving the development and popularization of intelligent applications across fields.

References

1. J. Wen, Z. Zhang, Y. Lan, et al., "A survey on federated learning: Challenges and applications," International Journal of Machine Learning and Cybernetics, vol. 14, no. 2, pp. 513-535, 2023.
2. H. Xin and R. Pan, "Self-attention-based modeling of multi-source metrics for performance trend prediction in cloud systems," Journal of Computer Technology and Software, vol. 4, no. 4, 2025 . doi: 10.5281/zenodo.15559874.
3. Y. Wang, "Entity-aware graph neural modeling for structured information extraction in the financial domain," Transactions on Computational and Scientific Methods, vol. 4, no. 9, 2024.doi: https://doi.org/10.5281/zenodo.15661653

 doi:10.20944/preprints202508.0797.v1

4. B. Fang and D. Gao, "Collaborative multi-agent reinforcement learning approach for elastic cloud resource scaling," arXiv preprint arXiv:2507.00550, 2025.
5. S. Banabilah, M. Aloqaily, E. Alsayed, et al., "Federated learning review: Fundamentals, enabling technologies, and future applications," Information Processing & Management, vol. 59, no. 6, 103061, 2022.
6. M. Wei, "Federated meta-learning for node-level failure detection in heterogeneous distributed systems," Journal of Computer Technology and Software, vol. 3, no. 8, 2024. doi: https://doi.org/10.5281/zenodo.15735387.
7. P. M. Mammen, "Federated learning: Opportunities and challenges," arXiv preprint arXiv:2101.05428, 2021.
8. E. T. M. Beltrán, M. Q. Pérez, P. M. S. Sánchez, et al., "Decentralized federated learning: Fundamentals, state of the art, frameworks, trends, and challenges," IEEE Communications Surveys & Tutorials, vol. 25, no. 4, pp. 2983-3013, 2023.
9. T. Tang, "A meta-learning framework for cross-service elastic scaling in cloud environments," Journal of Computer Technology and Software, vol. 3, no. 8, 2024. doi: https://doi.org/10.5281/zenodo.15653482
10. B. Liu, N. Lv, Y. Guo, et al., "Recent advances on federated learning: A systematic survey," Neurocomputing, 128019, 2024.
11. H. Guan, P. T. Yap, A. Bozoki, et al., "Federated learning for medical image analysis: A survey," Pattern Recognition, 110424, 2024.
12. R. Meng, H. Wang, Y. Sun, Q. Wu, L. Lian and R. Zhang, "Behavioral anomaly detection in distributed systems via federated contrastive learning," *arXiv preprint* arXiv:2506.19246, 2025.
13. L. Zhu, W. Cui, Y. Xing and Y. Wang, "Collaborative optimization in federated recommendation: Integrating user interests and differential privacy," *Journal of Computer Technology and Software*, vol. 3, no. 8, 2024. doi: 10.5281/zenodo.15653492.
14. Y. Lin and P. Xue, "Multi-task learning for macro-economic forecasting based on cross-domain data fusion," *Journal of Computer Technology and Software*, vol. 4, no. 6, 2025.doi: https://doi.org/10.5281/zenodo.15809661.
15. T. Yang, "Transferable load forecasting and scheduling via meta-learned task representations," *Journal of Computer Technology and Software*, vol. 3, no. 8, 2024. doi: 10.5281/zenodo.15735426.
16. X. Liu, Y. Qin, Q. Xu, Z. Liu, X. Guo and W. Xu, "Integrating knowledge graph reasoning with pretrained language models for structured anomaly detection," 2025. DOI:10.20944/preprints202505.1782.v1
17. M. Gong, "Modeling microservice access patterns with multi-head attention and service semantics," *Journal of Computer Technology and Software*, vol. 4, no. 6, 2025.DOI: 10.5281/zenodo.15809649.
18. S. Lyu, Y. Deng, G. Liu, Z. Qi and R. Wang, "Transferable modeling strategies for low-resource LLM tasks: A prompt and alignment-based," *arXiv preprint* arXiv:2507.00601, 2025.
19. Y. Xing, "Bootstrapped structural prompting for analogical reasoning in pretrained language models," *Transactions on Computational and Scientific Methods*, vol. 4, no. 11, 2024. DOI:https://doi.org/10.5281/zenodo.16750231
20. W. Zhang, Z. Xu, Y. Tian, Y. Wu, M. Wang and X. Meng, "Unified instruction encoding and gradient coordination for multi-task language models," 2025. DOI:10.20944/preprints202506.0582.v1
21. Y. Zhao, W. Zhang, Y. Cheng, Z. Xu, Y. Tian and Z. Wei, "Entity boundary detection in social texts using BiLSTM-CRF with integrated social features," 2025. DOI:10.20944/preprints202506.0582.v1.
22. L. Collins, H. Hassani, A. Mokhtari, et al., "FedAvg with fine tuning: Local updates lead to representation learning," Advances in Neural Information Processing Systems, vol. 35, pp. 10572-10586, 2022.
23. S. P. Karimireddy, S. Kale, M. Mohri, et al., "SCAFFOLD: Stochastic controlled averaging for federated learning," Proceedings of the International Conference on Machine Learning, PMLR, pp. 5132-5143, 2020.

 doi:10.20944/preprints202508.0797.v1

24. A. Fallah, A. Mokhtari, A. Ozdaglar, "Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach," Advances in Neural Information Processing Systems, vol. 33, pp. 3557-3568, 2020.
25. D. Zeng, S. Liang, X. Hu, et al., "FedLab: A flexible federated learning framework," Journal of Machine Learning Research, vol. 24, no. 100, pp. 1-7, 2023.