

Efficient and Scalable Cluster-Based Multi-Model Federated Learning Framework with Knowledge Distillation for Privacy-Preserving Healthcare Systems

S. Deeparani¹, P. Prema², V. Nagamalini³, Harini Sree K.⁴, Malathi R.⁵, Krihnapriya V.⁶

^{1,2,3} Assistant Professor/CSE, Sri Venkateswara college of Engineering, Tamil Nadu, India

^{4,5,6} UG Scholar, Dept. of CSE (ADS), Sri Venkateswara College of Engineering, Tamil Nadu, India

Emails: deeparanis@svce.ac.in¹, premap@svce.ac.in², nagamaliniv@svce.ac.in³, 2024ad0754@svce.ac.in⁴, 2024ad0327@svce.ac.in⁵, 2024ad0033@svce.ac.in⁶

Abstract

Federated Learning (FL) enables collaborative model training across distributed healthcare institutions without requiring the exchange of sensitive patient data. However, existing FL systems face challenges such as high communication overhead, data heterogeneity, and limited scalability. Additionally, the reliance on a single global model restricts the ability to effectively process multi-modal healthcare data. This paper proposes a cluster-based multi-model federated learning framework integrated with knowledge distillation. Clients are grouped based on data similarity, and specialized models are assigned to each cluster. Model outputs are fused at the central server, and knowledge distillation is used to compress multiple models into a lightweight global model. The framework is designed within an edge-cloud architecture to support scalable and practical deployment in real-world healthcare systems.

Keywords: Federated Learning, Healthcare Systems, Knowledge Distillation, Data Heterogeneity, Multi-Model Learning, Edge Computing

1. Introduction

The rapid growth of digital healthcare systems has led to an increase in distributed and sensitive medical data across hospitals and IoMT devices. While machine learning techniques can leverage this data for improved diagnostics and decision-making, privacy concerns and regulatory constraints limit centralized data sharing. Federated Learning (FL) addresses this challenge by enabling decentralized training, where clients share model updates instead of raw data (Briggs et al., 2020; Martínez Beltrán et al., 2023; Oh & Nadkarni, 2023). Despite its advantages, FL faces several limitations in healthcare applications. Data heterogeneity across institutions results in non-identically distributed (non-IID) data, reducing model performance (Kim et al., 2020; Ghosh et al., 2020). Communication overhead from frequent updates further limits scalability. Additionally, traditional FL approaches rely on a single global model, which is not suitable for handling diverse multi-modal healthcare data such as images, structured records, and time-series signals.

To address these issues, this paper proposes a cluster-based multi-model federated learning framework that

integrates clustering, model specialization, fusion, and knowledge distillation.

2. Related Work

2.1. Federated Learning Challenges in Healthcare Systems

Federated Learning (FL) has evolved significantly as a decentralized learning paradigm, with recent research focusing on improving its efficiency, scalability, and adaptability in real-world environments (Martínez Beltrán et al., 2023; Mammen, 2021). While early approaches provide a baseline for distributed model aggregation, they often assume relatively homogeneous data distributions, which is not realistic in practical healthcare settings (Mammen, 2021). Healthcare data is often highly heterogeneous and unbalanced across institutions, which further complicates model training and generalization in federated settings (Chaddad et al., 2023; Chung et al., 2026). Non-IID data distributions significantly affect model convergence and often require additional communication rounds to achieve acceptable performance in federated learning systems (Chung et al., 2026). These challenges highlight the

need for improved frameworks that can handle distributional variability across clients. Experimental studies have also shown that the performance of FedAvg-based federated[1] learning is highly dependent on the homogeneity of data distributions, particularly in medical imaging applications, where heterogeneous datasets can lead to unstable training behavior (Kovalev & Karpenka, 2022).

2.2. Clustered and Hierarchical Federated Learning Approaches

To address non-IID data issues, clustered federated learning has been widely explored. Early works introduce hierarchical clustering[2] of local updates to improve training stability under heterogeneous data distributions (Briggs et al., 2020). Dynamic clustering techniques further enhance adaptability by allowing clients to change cluster assignments during training based on update similarity (Kim et al., 2020). Ghosh et al. proposed an efficient clustered federated learning framework that groups clients with similar data distributions to improve convergence and accuracy (Ghosh et al., 2020). Similarly[3], recent studies have explored optimized clustering strategies that enhance communication efficiency and robustness in heterogeneous environments (Zhang et al., 2025). Survey studies also confirm that clustering-based FL[4] significantly improves personalization and reduces statistical heterogeneity (Liu et al., 2025). In addition, edge and resource-efficient clustered FL frameworks have been proposed for real-world deployment scenarios, particularly in constrained environments such as IoT and Industry 4.0 systems (Takele & Villányi, 2025).

2.3. Communication Efficiency and System Scalability

One major direction of research focuses on improving communication efficiency in FL systems. Frequent exchange of model updates between clients and the central server introduces significant overhead, especially in large-scale deployments. To address this, techniques such[5] as model compression, gradient sparsification, and selective update strategies have been proposed. In addition, device clustering and communication optimization methods have been explored to reduce overhead while maintaining model performance (Wei et al., 2025). These approaches improve scalability by

reducing redundant communication and enabling efficient coordination among heterogeneous devices. Hierarchical and edge-based federated learning architectures have also been explored to enhance scalability and reduce[6] latency (Briggs et al., 2020; Takele & Villányi, 2025). However, interoperability between heterogeneous healthcare systems remains a significant challenge, as data is often stored in incompatible formats across institutions, limiting effective data utilization (Akhmetov et al., 2025).

2.4. Personalized, Multimodal, and Knowledge-Driven Federated Learning

Another important direction is personalized federated learning, where models are adapted to individual clients instead of relying entirely on a shared global model. These approaches aim to improve performance under heterogeneous data distributions by allowing local customization while still benefiting from collaborative training [6]. Recent work has also investigated multi-task and multi-model federated learning, where different models are trained to handle diverse data types or tasks. This is particularly relevant in healthcare environments involving multimodal data such as medical images, electronic health records, and time-series signals from wearable devices (Wang et al., 2025). Knowledge distillation techniques have also been explored to reduce model complexity and communication overhead. Instead of sharing full model parameters, these approaches transfer knowledge in a compressed form, enabling lightweight model deployment in resource-constrained environments.

2.5. Research Gap

Despite these advancements, most existing works address clustered learning, communication efficiency, multimodal learning, and model compression independently. There remains a lack of unified frameworks[7] that integrate clustering, multi-model learning, and knowledge distillation within a scalable federated learning architecture. Clustered FL approaches such as (Ghosh et al., 2020) and (Zhang et al., 2025) improve heterogeneity handling, while surveys like (Liu et al., 2025) highlight scalability challenges. However, multimodal healthcare FL systems (Chaddad et al., 2023; Wang et al., 2025) and communication-optimized frameworks (Wei et al., 2025) are often

developed separately, limiting their combined effectiveness in real-world deployment scenarios. The proposed method aims to bridge this gap by integrating clustering, multi-model learning, model fusion, and knowledge distillation into a unified federated learning framework for healthcare environments[8].

3. Problem Statement

Existing federated learning approaches in healthcare face several fundamental limitations when deployed in real-world heterogeneous environments. Although federated learning enables collaborative training without sharing raw data, performance degradation occurs due to highly non-IID and imbalanced data distributions across healthcare institutions. Studies have also demonstrated that FedAvg exhibits unstable behavior when applied to heterogeneous medical imaging datasets, highlighting its limitations in real-world healthcare scenarios (Kovalev & Karpenka, 2022). Most existing methods either rely on a single global model or address data heterogeneity using clustering techniques alone. However, clustering-based approaches still assume a uniform model architecture across all clients, which limits their ability to handle multimodal healthcare data such as medical images, electronic health records, and time-series signals. Furthermore, existing personalized and clustered federated learning methods often overlook the need for model specialization across different data modalities, leading to suboptimal feature representation and reduced predictive performance. In addition, communication overhead and system scalability remain critical challenges, particularly in large-scale deployments involving resource-constrained edge devices. Finally, most existing approaches do not provide an efficient mechanism to compress multiple specialized models into a unified lightweight model suitable for real-world deployment.[9]Therefore, there is a need for a unified federated learning framework that integrates client clustering, multi-model learning, model fusion, and knowledge distillation to effectively handle data heterogeneity, improve learning efficiency, and enable scalable deployment in healthcare environments[10].

4. Proposed Methodology

The proposed framework introduces a cluster-based

multi-model federated learning system enhanced with knowledge distillation for efficient and privacy-preserving healthcare analytics. The system is designed to handle[11] heterogeneous and non-Independent and Identically Distributed (non-IID) data across distributed healthcare clients. Traditional aggregation methods such as parameter averaging require homogeneous model architectures and are not suitable for heterogeneous federated settings (Lin et al., 2020).

4.1.Client Clustering

Initially, clients are grouped into clusters based on similarity measures derived from their local data characteristics. Clustered federated learning groups clients with similar[12] data distributions, enabling more effective model training within each group and reducing the impact of non-IID data distributions (Zhang et al., 2025; Ye et al., 2023). This clustering improves convergence stability and training efficiency by ensuring that updates within each cluster are more consistent (Liu et al., 2025). Shown in Figure 1 Cluster-Based Federated Learning Flow Architecture showing client grouping based on data similarity and the formation of clusters for distributed training[13].

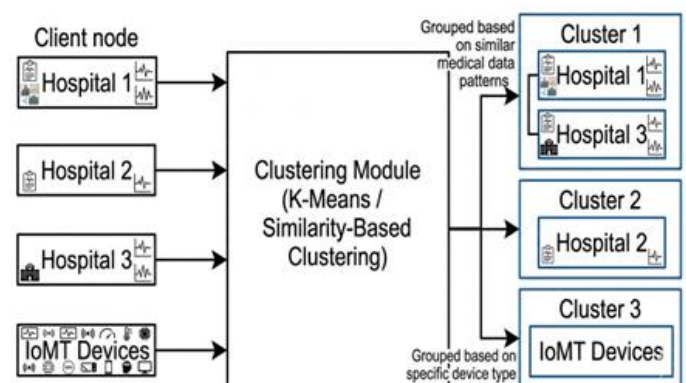


Figure 1 Cluster-Based Federated Learning Flow Architecture showing client grouping based on data similarity and the formation of clusters for distributed training

4.2.Multi-Model Learning and Local Training

Within each cluster, specialized models are assigned based on the type of healthcare data. Convolutional Neural Networks (CNNs) are used for medical image

data; Long Short-Term Memory (LSTM) or Transformer models are used for time-series data, and tree-based models are applied to structured clinical records[14]. Each client performs local training using its private dataset, and only model updates are transmitted to the central server. This ensures data privacy while enabling collaborative learning across distributed healthcare institutions.

4.3. Server Aggregation and Model Fusion

At the server level, updates from clients within each cluster are aggregated using a federated aggregation mechanism to form cluster-specific global models. Since each cluster captures different data distributions, their outputs are complementary[15]. A fusion strategy is then applied to combine predictions from multiple cluster-specific models using weighted aggregation. The weights are assigned based on model performance or confidence scores, ensuring that more reliable models contribute more significantly to the final prediction.

4.4. Knowledge Distillation and Model Compression

To reduce computational complexity and enable deployment in resource-constrained environments, knowledge distillation is applied to compress multiple cluster-specific models into a single lightweight global model (Salman et al., 2025). In this process, the ensemble of cluster models acts as the teacher model, while the global model acts as the student. The teacher generates soft probability outputs that contain richer information than hard labels. The student model is trained to mimic these outputs by minimizing the divergence between teacher and student predictions, thereby effectively transferring knowledge in the federated learning setting (Salman et al., 2025). This enables efficient knowledge transfer while maintaining high predictive performance in a compact model suitable for edge-based healthcare systems[16 – 20].

4.5. Edge–Cloud Architecture Overview

The overall system operates within an edge–cloud architecture. Edge devices perform local training and data processing, while the cloud server handles clustering, aggregation, fusion, and model distillation. This ensures scalability, reduced communication overhead, and efficient deployment in real-world healthcare environments[21].

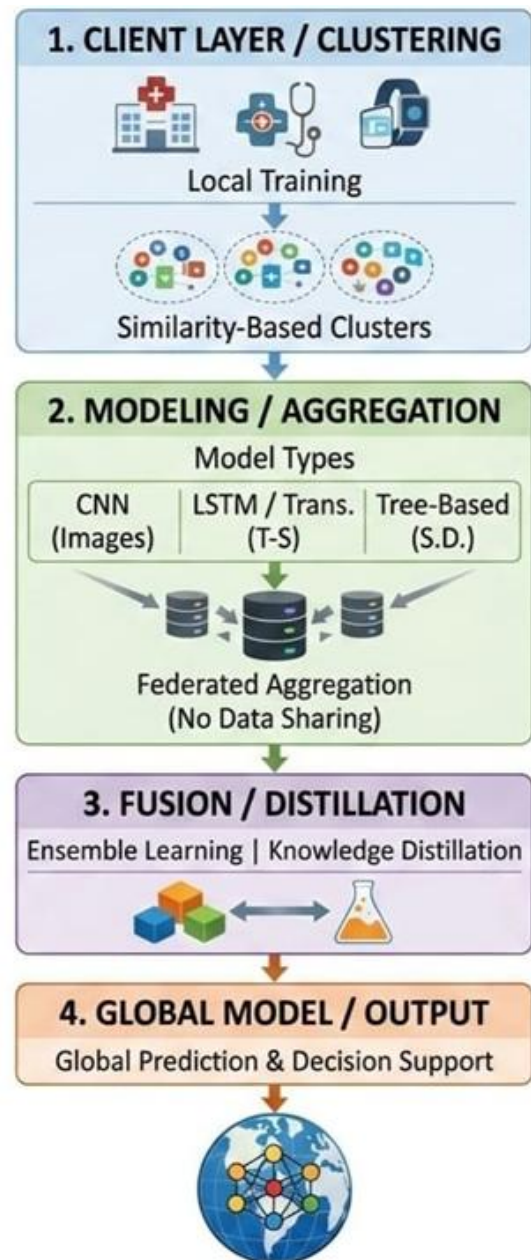


Figure 2 Architecture of the Proposed Cluster-Based Multi-Model Federated Learning Framework

5. Algorithm

Algorithm 1: Cluster-Based Multi-Model Federated Learning

- Initialize global model parameters
- Collect client metadata and perform similarity-based clustering (Wei et al., 2025)
- Assign model type to each cluster based on

data modality [22]

- For each communication round:
- Distribute cluster-specific models to clients
- Perform local training at each client
- Send model updates to central server
- Aggregate updates within clusters
- Apply model fusion across clusters
- Perform knowledge distillation to obtain global lightweight MODEL[23]
- Update global model and repeat until convergence

The proposed algorithm describes the overall flow of cluster-based multi-model federated learning framework. First, global model parameters are set and similar clients are linked into clusters by the similarity in their data characteristics. The benefit of this clustering step is to deal with non-IID data and to train within more homogenous groups, which is explored in various approaches of clustered federated learning (Chung et al., 2026). Cluster-specific models are given out to clients for each communication round. All clients train locally with their own data, and only updates to the model are sent to the central server, guaranteeing data protection. This is done in the traditional federated learning paradigm of making decentralized training and centralized aggregation (Zhang et al., 2025; Zhang et al., 2024). The server then combines updates in each cluster to produce cluster-level models. The mechanism of fusion allows combining the results of various specialized models to draw knowledge from the clusters. Finally, knowledge distillation is carried out to reduce the ensemble of models to a lightweight global model that is then redistributed for the next round of training. This iterative process is continued until the model converges[24].

6. System Design and Architecture

This system is built of layers, namely client, clustering, model, aggregation, fusion, and output layers. The client layer consists of healthcare organizations and IoMT devices where the local training is conducted. The clustering layer aggregates clients to deal with data heterogeneity (Wang et al., 2025). There are special models for different types of data in the model layer. The aggregation layer maintains secure aggregation of updates without

sharing raw data. The fusion layer averages the answers from a variety of models to increase performance. The final output layer provides a prediction, e.g., disease diagnosis or risk assessment.

Table 1 Model Assignment

Data Type	Model Used
Medical Images	CNN
Time-Series Data	LSTM / Transformer
Structured Data	Tree-based Models

The proposed approach is hybrid in nature, which will address the challenges involved in traditional approaches by combining both clustering and multi-model learning algorithms, which in turn will address the challenges of multi-modal and heterogeneous data and reduce communication overhead[25].

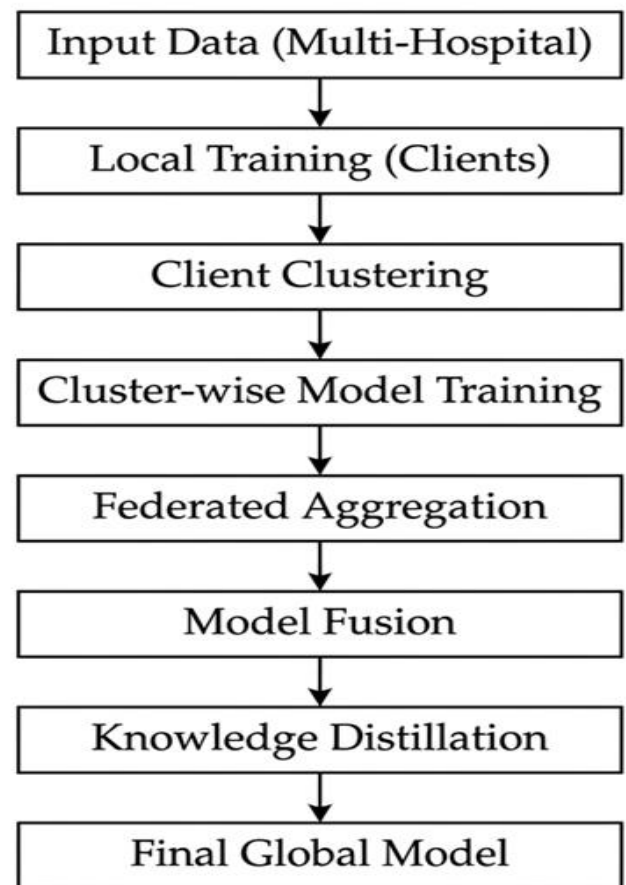


Figure 3 Workflow of the proposed cluster-based multi-model federated learning framework for healthcare systems

7. Comparison With Existing Methods

Table 2 Comparison of Approaches

Method	Handles non-IID	Multi-Model	Communication Efficient
FedAvg	No	No	Moderate
Clustered FL	Yes	No	Moderate
Proposed Method	Yes	Yes	High

8. Experimental Setup

The simulated federated learning environment represents multiple healthcare institutions in which the proposed framework is evaluated. Real-world clinical data information is not used for this; therefore, a federated environment is synthesized so as to simulate data heterogeneity on different clients. It is assumed that each client has locally stored data with non-Independent and Identically Distributed (non-IID) nature. To resemble real-world variations in healthcare data, the data distribution is artificially partitioned using strategies like Dirichlet-based partitioning used in federated learning experiments (Jimenez-Gutierrez et al., 2026) to yield variations in class imbalance in the dataset, as well as feature distributions. These partitioning methods are commonly employed for scenarios with different degrees of non-IID data at the clients, allowing the study of the federated learning performance to be assessed under the different data partitionings (Jimenez-Gutierrez et al., 2026). The clients receive the data modalities, including medical images, time-series signals, and structured clinical records, each representing the various healthcare environments. The federated learning process is simulated by having multiple rounds of communication between clients and a central server. Standard machine learning metrics (accuracy, precision, recall and F1-score) are used to describe the performance evaluation of the described method. Furthermore, communication overhead and convergence behavior are taken into account to analyze the efficiency of a system. Due to lack of real-world deployment, the results reported are qualitative in nature.

9. Results And Discussion

The proposed framework is tested in a simulated federated learning environment that simulates the heterogeneous healthcare environment. The outcomes are presented with regard to the model performance, communication efficiency, and scalability of the system. Many researchers have noticed that the performance of federated learning models tends to degrade as the distribution of data changes from IID to non-IID, and in highly skewed scenarios, the models can become unstable and less accurate (Domini et al., 2025). The clustering mechanism will group clients into clusters where the data distributions are similar and thereby enhance the overall model performance. This eliminates the effect of non-IID data and helps to obtain more robust convergence than conventional federated learning methods (Takele & Villányi, 2025). Table II demonstrates that clustered federated learning and the proposed approach are superior to standard FedAvg, which is effective in handling data heterogeneity. The fusion of various specialized models facilitates the effective extraction of features from the various types of healthcare data, such as medical images, time-series signals, and clinical records. This multi-model approach is more effective in predicting than a single global model-based approach. Besides, the fusion strategy would enhance the robustness of prediction in the sense that it would fuse the output of several cluster-specific models, hence reducing the variance and improving the overall quality of prediction. Moreover, knowledge distillation further compresses the models into a lightweight global model that is suitable for deployment in the resource-constrained environment. Moreover, clustering minimizes redundant communications between different clients, which can help to enhance the efficiency of communication and lower the overhead on the network. The edge-cloud architecture also provides greater scalability and capability for distributed processing to support real-time healthcare applications. In general, the proposed framework provides better adaptability, efficiency, and scalability in the processing of heterogeneous federated learning environments.

Conclusion And Future Work

In this paper, a multi-model federated learning

framework integrated by knowledge distillation is introduced for the application of healthcare. The designed approach covers a variety of problems originating from data heterogeneity, communication overhead, and scaling in distributed learning settings. The framework features client clustering, specialized model training, prediction fusion, and knowledge distillation, ensuring an efficient and scalable approach to heterogeneous healthcare data analysis. The convergence of models in clustered federated learning and the lightweight deployment in edge-based systems in knowledge distillation are achieved. Future work will include testing the framework with real-life clinical data and adding more advanced security measures, like differentially private methods or secure multi-party computation. Further, adaptive clustering schemes and dynamic model selection schemes may be investigated for further enhancing system performance in a changing healthcare environment.

Acknowledgements

We would like to express our sincere gratitude to Sri Venkateswara College of Engineering for their continuous support and encouragement in carrying out this research on the Efficient and Scalable Cluster-Based Multi-Model Federated Learning Framework with Knowledge Distillation for Privacy-Preserving Healthcare Systems. We would also like to acknowledge the valuable contributions of the research community in the fields of federated learning, healthcare analytics, and distributed artificial intelligence, whose published works and publicly available resources greatly supported the conceptual development of this study. In addition, we extend our appreciation to all faculty members, colleagues, and peers whose insightful discussions and constructive suggestions helped refine the ideas and methodology presented in this work. Lastly, we thank everyone who directly or indirectly contributed to the successful completion of this research.

References

- [1]. C. Briggs, Z. Fan, and P. Andras, "Federated Learning with Hierarchical Clustering of Local Updates to Improve Training on Non-IID Data," arXiv preprint arXiv:2004.11791, 2020.
- [2]. Y. Kim, E. Al Hakim, J. Haraldson, H. Eriksson, J. M. B. da Silva Jr., and C. Fischione, "Dynamic Clustering in Federated Learning," arXiv preprint arXiv:2012.03788, 2020.
- [3]. A. Ghosh, J. Chung, D. Yin, and K. Ramchandran, "An Efficient Framework for Clustered Federated Learning," in Advances in Neural Information Processing Systems (NeurIPS), 2020.
- [4]. E. T. Martínez Beltrán, M. Quiles Pérez, P. M. Sánchez Sánchez, S. López Bernal, G. Bovet, M. Gil Pérez, G. Martínez Pérez, and A. Huertas Celdrán, "Decentralized Federated Learning: Fundamentals, State of the Art, Frameworks, Trends, and Challenges," IEEE Communications Surveys & Tutorials, 2023.
- [5]. P. M. Mammen, "Federated Learning: Opportunities and Challenges," in Proceedings of the ACM Conference, 2021.
- [6]. A. Chaddad, Y. Wu, and C. Desrosiers, "Federated Learning for Healthcare Applications," 2023.
- [7]. A. Akhmetov, Z. Latif, B. Tyler, and A. Yazici, "Enhancing Healthcare Data Privacy and Interoperability with Federated Learning," PeerJ Computer Science, vol. 11, e2870, 2025.
- [8]. W.-C. Chung, C.-A. Lo, Y.-H. Lin, Z.-H. Chen, and C.-L. Hung, "Decentralized Federated Learning with Non-IID Data: Challenges, Trends, and Future Opportunities," ACM Computing Surveys, vol. 58, no. 8, article 192, 2026.
- [9]. Y. Zhang, H. Chen, Z. Lin, Z. Chen, and J. Zhao, "LCFed: An Efficient Clustered Federated Learning Framework for Heterogeneous Data," in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2025.
- [10]. D. Wei, X. Xu, S. Mao, and M. Chen, "Optimizing Communication and Device Clustering for Clustered Federated Learning with Differential Privacy," IEEE Transactions on Mobile Computing, 2025.

- [11]. E. Liu, W. Yang, Y. Gu, W. Long, Szabó István, and L. Jiang, "A Survey of Clustering Federated Learning in Heterogeneous Data Scenarios," 2025.
- [12]. X. Wang, R. Zhou, H. Xie, X. Tang, L. He, and Carl Yang, "ClusMFL: A Cluster-Enhanced Framework for Modality-Incomplete Multimodal Federated Learning in Brain Imaging Analysis," arXiv preprint arXiv:2502.12180, 2025.
- [13]. Atallo Kassaw Takele and Balázs Villányi, "Resource-Efficient Clustered Federated Learning Framework for Industry 4.0 Edge Devices," AI, vol. 6, no. 2, p. 30, 2025.
- [14]. Yuxin Zhang, Haoyu Chen, Zheng Lin, Zhe Chen, and Jin Zhao, "FedAC: An Adaptive Clustered Federated Learning Framework for Heterogeneous Data," arXiv preprint arXiv:2403.16460, 2024.
- [15]. V. Kovalev and D. Karpenka, "The Efficiency of FedAVG Federated Learning in Training Generative Image Models," in Proceedings of the International Conference on Biomedical Image Analysis, 2022.
- [16]. D. M. Jimenez-Gutierrez, M. Hassanzadeh, D. Solans, M. Elbamby, N. Kourtellis, A. Anagnostopoulos, I. Chatzigiannakis, and A. Vitaletti, "Clust-PSI-PFL: A Population Stability Index Approach for Clustered Non-IID Personalized Federated Learning," arXiv preprint arXiv:2512.20363, 2026.
- [17]. D. Domini, G. Aguzzi, and M. Viroli, "ProFed: a Benchmark for Proximity-based Non-IID Federated Learning," arXiv preprint arXiv:2503.20618, 2025.
- [18]. H. Salman, C. Zaki, N. Charara, S. Guehis, J.-F. Pradat-Peyre, and A. Nasser, "Knowledge distillation in federated learning: a comprehensive survey," Discover Computing, vol. 28, Art. no. 145, 2025.
- [19]. M. Ye, X. Fang, B. Du, P. C. Yuen, and D. Tao, "Heterogeneous Federated Learning: State-of-the-art and Research Challenges," ACM Computing Surveys, vol. 56, no. 3, Article 79, Oct. 2023.
- [20]. T. Lin, L. Kong, S. U. Stich, and M. Jaggi, "Ensemble Distillation for Robust Model Fusion in Federated Learning," in Advances in Neural Information Processing Systems (NeurIPS), 2020.
- [21]. W. Oh and G. N. Nadkarni, "Federated Learning in Health Care Using Structured Medical Data," Advances in Kidney Disease and Health, vol. 30, no. 1, pp. 4–16, 2023.