Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Personalized Federated Learning Under Severe Statistical Heterogeneity: A Multi-Dataset Analysis of Accuracy, Tail Performance, and Client Fairness

Md Shahanur Islam Shagor

smshagor.ru@gmail.com

Voronezh State University of Forestry and Technologies https://orcid.org/0009-0003-7730-3202

Research Article

Keywords: client fairness, federated learning, non-IID data, personalization, personalized federated learning, statistical heterogeneity

Posted Date: September 21st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11073326/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Personalized Federated Learning Under Severe Statistical Heterogeneity: A Multi-Dataset Analysis of Accuracy, Tail Performance, and Client Fairness

Independent Researcher, Department of Computer Science and Microelectronics Engineering,
Voronezh State University of Forestry and Technologies, Voronezh, Russia
Corresponding author: smshagor.ru@gmail.com

Abstract—Personalized federated learning (PFL) is intended to mitigate the degradation caused by statistically heterogeneous client data, yet average accuracy alone cannot establish whether personalization improves the clients that are most poorly served by a shared model. This paper develops a client-centric analytical and multi-dataset evaluation framework for studying utility, lower-tail performance, and performance-distribution fairness under severe non-independent and identically distributed data. The framework separates nominal partition parameters from realized heterogeneity, uses matched client partitions and random seeds for paired algorithm comparisons, and jointly evaluates global accuracy, mean and median client accuracy, tenth-percentile and worst-client accuracy, dispersion, and Jain's fairness index. Three analytical results motivate the design. For symmetric Dirichlet label allocation, the expected client concentration is an explicit monotone function of the Dirichlet parameter, while realized partitions remain random; Jain's index is exactly the inverse of one plus the squared coefficient of variation of nonnegative client accuracies, demonstrating that fairness must be interpreted together with absolute utility; and matched-seed comparisons reduce comparison variance whenever paired outcomes are positively correlated. The evaluation protocol covers MNIST, FashionMNIST, CIFAR-10, and CIFAR-100 and contrasts FedAvg, FedProx, SCAFFOLD, Ditto, Per-FedAvg, and FedSAM across IID, Dirichlet label skew, pathological class skew, and quantity skew. The resulting methodology provides a rigorous basis for determining when personalization improves not only average performance but also the lower tail of the client population.

Index Terms—client fairness, federated learning, non-IID data, personalization, personalized federated learning, statistical heterogeneity.

I. INTRODUCTION

Federated learning enables a population of clients to optimize a model without centralizing their raw training data. Its statistical difficulty is equally fundamental: clients rarely sample from the same distribution. Differences in label prevalence, local sample volume, sensing conditions, user behavior, and task context produce conflicting local objectives and can substantially degrade both convergence and client-level generalization. Recent empirical studies have therefore treated heterogeneity as a primary determinant of federated learning behavior rather than as a secondary implementation detail [1], [5], [6].

Personalized federated learning addresses this problem by allowing the model used by a client to differ, in a controlled manner, from one shared global solution. Current PFL approaches include adaptive aggregation, layer-wise collaboration, regularized client-specific objectives, meta-learning, representation sharing, and sharpness-aware local optimization [2]–[4], [11], [14]. The motivation is straightforward: when client distributions differ substantially, a single global optimum may represent an acceptable population average while remaining systematically unsuitable for individual clients.

Personalization, however, is not synonymous with fairness. A method can increase global or average client accuracy while leaving the weakest clients unchanged or even worse off. Conversely, a procedure can make client accuracies more equal while lowering their absolute level. Recent work on heterogeneous and personalized federated learning has consequently emphasized fairness, lower-tail behavior, and trade-offs among accuracy, privacy, communication, and client disparity [7], [8], [10], [12], [16]. These observations make the distribution of client outcomes, rather than only its mean, a central object of evaluation.

A second difficulty is experimental comparability. PFL studies often differ in client count, partition generator, random seeds, model architecture, local optimization budget, and evaluation protocol. A systematic PFL benchmark showed that method rankings can change across datasets and experimental settings and that comparatively simple federated baselines followed by local adaptation may remain competitive in some regimes [5]. When competing algorithms are evaluated on independently sampled non-IID partitions, part of an apparent algorithmic difference can therefore be caused by differences in realized heterogeneity.

This paper develops a client-centric analytical and experimental methodology for evaluating PFL under severe statistical heterogeneity. It does not define fairness through a single scalar. Instead, it combines global utility, central client utility, lower-tail utility, dispersion, and Jain's index under matched partitions. The framework also treats the partition generator itself as an experimental variable by distinguishing IID allocation, Dirichlet label skew, pathological class restriction, and quantity skew.

The main contributions are as follows. First, the paper formalizes why nominal Dirichlet parameters do not uniquely determine realized heterogeneity and derives a closed-form concentration measure that changes monotonically with the Dirichlet parameter. Second, it derives an exact relationship between Jain's fairness index and the coefficient of variation of client accuracy, clarifying why fairness and accuracy must be interpreted jointly. Third, it establishes the variance advantage of matched-seed algorithm comparisons under positive paired correlation. Fourth, it specifies a reproducible multi-dataset evaluation design for six representative global and personalized federated methods, with lower-tail client metrics and corrected paired statistical comparisons.

II. RELATED WORK

A. Personalized Federated Learning

Tan et al. organized PFL around approaches that adapt a global model, learn client-specific models jointly, cluster clients, or exploit meta-learning and transfer [1]. pFedLA introduced

layer-wise aggregation to exploit similarity at finer granularity than whole-model weighting [3], while FedALA adaptively combines global and local parameters before local optimization [4]. These approaches reflect a broader change in perspective: the server model is not necessarily the final model for every participant but a transferable source of information whose relevance varies across clients.

Sharpness-aware optimization provides a complementary route to robustness under distribution shift. FedSAM introduced sharpness-aware local training for federated learning [2], and later work examined local estimates of global perturbations and personalized sharpness-aware training [11], [13]. Recent 2025–2026 PFL methods further combine representation sharing, adaptive aggregation, communication efficiency, or composite feature extraction to cope with increasingly heterogeneous client populations [12], [14], [17].

Comparative methodology remains important because additional personalization machinery does not guarantee consistent gains. Matsuda et al. demonstrated substantial sensitivity of PFL rankings to datasets and settings [5]. The present work therefore emphasizes matched partitions and distributional client metrics rather than proposing a seventh optimization method.

B. Fairness Under Client Heterogeneity

Fairness in federated learning has several meanings, including selection fairness, incentive fairness, demographic group fairness, and equality of model performance across clients. This paper studies the last of these: performance-distribution fairness. Heterogeneity-aware fair federated learning has shown that conflicting local objectives can generate unequal client outcomes even when the aggregate objective converges [7], while broader surveys identify fairness as a persistent requirement for trustworthy federated systems [8].

The relationship between personalization and fairness is particularly nuanced. Zhao et al. studied convergence, privacy, and fairness jointly in personalized federated learning [10], and FedCoSR combined shareable representations with adaptive local aggregation to address accuracy and fairness under label heterogeneity [12]. FedGraph-Fair extended this direction by combining personalization with graph-based collaboration and robust objectives for worst-off clients [16]. These studies reinforce the need to report lower-tail client behavior instead of interpreting mean accuracy as a fairness surrogate.

The present methodology differs from methods that optimize a fairness regularizer directly. Its purpose is diagnostic: it asks whether a personalization mechanism improves the center and the lower tail of the client-performance distribution under the same realized data partition and training conditions.

III. PROBLEM FORMULATION

A. Federated Objective

Consider K clients. Client k holds D_k with n_k samples and local objective $F_k(w)$. A standard sample-weighted federated objective is

$$\min_w \; F(w) = \Sigma_k \, p_k F_k(w), \qquad p_k = n_k \,/\, \Sigma_j \, n_j \, .$$

When $P_k(X,Y)$ differs across clients, the minimizer of F can be far from the minimizer preferred by an individual client. A personalized method instead returns client-dependent parameters v_k, while preserving some form of shared information transfer.

B. Client Utility and Lower-Tail Metrics

Let $a_k \in [0,1]$ denote held-out accuracy for client k. The evaluation vector $a=(a_1,\ldots,aK)$ is summarized by the population mean μ, median, tenth percentile $q_{0.1}$, minimum a_min, and population dispersion σ.

$$\mu = (1/K)\Sigma_k a_k, \qquad \sigma^2 = (1/K)\Sigma_k (a_k - \mu)^2.$$

The lower tail is represented by $q_{0.1}$ and a_min. These metrics answer a different question from μ: whether a method improves clients that are least well served by the current model.

C. Jain Fairness

For nonnegative client accuracies, Jain's fairness index is

$$J(a) = (\Sigma_k a_k)^2 \,/\, [K\Sigma_k a_k^2], \qquad 1/K \le J \le 1.$$

J is invariant to common positive scaling. It therefore quantifies relative equality, not absolute predictive quality. A valid fairness interpretation must be paired with at least one absolute utility measure.

IV. LEARNING METHODS CONSIDERED

A. Shared-Model Baselines

FedAvg aggregates locally trained client updates into a shared global model. FedProx adds a proximal penalty that discourages local parameters from moving too far from the current global state:

$$\min_w \; F_k(w) + (\mu_p/2)||w - w_t||_2^2.$$

SCAFFOLD uses server and client control variates to reduce client drift caused by heterogeneous objectives. These methods provide shared-model baselines with increasing mechanisms for controlling heterogeneity.

B. Personalized Objectives

Ditto learns a personalized model v_k regularized toward a shared model w:

$$\min_{v_k} \; F_k(v_k) + (\lambda/2)||v_k - w||_2^2.$$

Per-FedAvg interprets the shared parameters as an initialization that should adapt rapidly to a client. With one inner adaptation step,

$$w'_k = w - \alpha \nabla F_k(w),$$

and the meta-objective is evaluated after adaptation. Ditto and Per-FedAvg thus represent two distinct meanings of personalization: regularized client-specific optimization and rapid client-specific adaptation.

C. Sharpness-Aware Federated Training

FedSAM applies a local sharpness-aware perturbation. For local gradient g, a representative perturbation is

$$\varepsilon = \rho g \,/\, (||g||_2 + \tau),$$

followed by gradient evaluation in the perturbed neighborhood. It is included because heterogeneity affects not only client drift but also the local loss geometry and generalization of aggregated parameters [2], [13].

V. HETEROGENEITY MODELS

A. Dirichlet Label Skew

For class c, client proportions are sampled from a symmetric Dirichlet distribution

$$(\pi_1 c,\ldots,\pi Kc) \sim \mathrm{Dirichlet}(\alpha,\ldots,\alpha).$$

The parameter α controls expected concentration but does not uniquely determine a realized partition. The evaluation therefore stores and reuses the same realized allocation for all algorithms compared within a matched cell.

B. Pathological and Quantity Skew

Pathological class skew allocates each client only a restricted subset of labels, producing explicit support mismatch. Quantity skew instead changes the number of samples owned by each client. A convenient unnormalized quantity model is $z_k \sim \text{LogNormal}(0,\sigma_q^2)$, followed by normalized allocation shares $q_k=z_k/\Sigma_j z_j$. These mechanisms are evaluated separately because label concentration and sample scarcity create different optimization difficulties.

VI. ANALYTICAL RESULTS AND DESIGN IMPLICATIONS

A. Dirichlet Concentration Is Monotone in α

For the symmetric K-dimensional Dirichlet distribution above, the first two moments of the allocation to client k are

$$E[\pi_k c]=1/K, \quad \text{Var}(\pi_k c)=(K-1)/[K^2(K\alpha+1)].$$

A convenient class-wise concentration index is $C_c=\Sigma_k \pi_k c^2$. Using the second moment of a Dirichlet component gives

$$E[C_c] = (\alpha+1)/(K\alpha+1).$$

The derivative with respect to α is negative for K>1:

$$dE[C_c]/d\alpha = (1-K)/(K\alpha+1)^2 < 0.$$

Thus smaller α produces greater expected class concentration, but C_c remains random at any finite α. Two experiments with the same α can therefore realize materially different class allocations. This formally supports matched-partition comparisons and reporting realized heterogeneity rather than treating α as a complete description of a partition.

B. Jain Fairness Equals an Inverse Relative-Dispersion Measure

Because $(1/K)\Sigma_k a_k^2 = \mu^2+\sigma^2$, Jain's index can be rewritten as

$$J = \mu^2/(\mu^2+\sigma^2) = 1/[1+(\sigma/\mu)^2], \quad \mu>0.$$

Hence J is exactly determined by the coefficient of variation of client accuracy. Increasing J means reducing relative dispersion; it does not imply that μ increased. Moreover, for any c>0, J(ca)=J(a). This proves that Jain fairness should never be reported without absolute accuracy or lower-tail utility. A method can improve equality by compressing all client accuracies toward a lower common value.

C. Quantity-Skew Severity Has a Monotone Dispersion Scale

For the unnormalized log-normal quantity variable z_k, the coefficient of variation is

$$CV(z) = \sqrt{[\exp(\sigma_q^2) - 1]}.$$

The expression is strictly increasing for $\sigma_q>0$. Consequently, σ_q has a direct interpretation as a monotone control on pre-normalization client-size dispersion. After normalization, client shares are dependent because they sum to one, but increasing σ_q still produces progressively more unequal candidate allocations in expectation.

D. Matched Comparisons Reduce Variance Under Positive Pairing

Let X and Y denote the seed-level metric produced by methods A and B on the same realized partition and random seed. For paired difference D=X−Y,

$$\text{Var}(D)=\text{Var}(X)+\text{Var}(Y)-2\text{Cov}(X,Y).$$

If the common partition and seed induce positive covariance, the paired comparison has lower variance than an unpaired comparison that discards this covariance. The implication is practical: algorithm comparisons should preserve partition identity and seed identity whenever possible. The independent observational unit remains the matched experiment seed, not a communication round.

E. Interpretation Rule for Personalized Fairness

The preceding identities motivate a two-axis interpretation. A personalization method provides a stronger fairness result only when a reduction in relative dispersion or an increase in lower-tail accuracy is accompanied by non-decreasing absolute utility under the same evaluation protocol. This avoids labeling uniformly degraded client predictions as a fairness improvement.

VII. MULTI-DATASET EVALUATION PROTOCOL

A. Datasets

The protocol uses MNIST, FashionMNIST, CIFAR-10, and CIFAR-100. The official training split is partitioned among clients. The official test split remains disjoint from optimization and is deterministically divided into non-overlapping held-out client subsets for client-level evaluation.

Dataset	Train/Test	Classes	Role
MNIST	60,000 / 10,000	10	Low-complexity reference
FashionMNIST	60,000 / 10,000	10	Moderate grayscale shift
CIFAR-10	50,000 / 10,000	10	Natural-image benchmark
CIFAR-100	50,000 / 10,000	100	High class diversity

TABLE I. MULTI-DATASET EVALUATION SET.

B. Matched Experiment Matrix

The primary design uses K=20 clients and five independent experiment seeds {11, 23, 37, 53, 71}. Every method in a comparison receives the same realized partition. Dirichlet experiments use $\alpha\in\{1.0,0.3,0.1,0.05\}$; IID, pathological class skew, and quantity-skew conditions provide complementary controls.

Dimension	Values
Methods	FedAvg; FedProx; SCAFFOLD; Ditto; Per-FedAvg; FedSAM
Clients	20
Seeds	11, 23, 37, 53, 71
Label skew	IID; Dirichlet α=1.0, 0.3, 0.1, 0.05
Other skew	Pathological class restriction; quantity skew
Primary metrics	Global, mean client, median, p10, worst-client, SD, Jain J

TABLE II. MATCHED EXPERIMENT DESIGN.

C. Evaluation Metrics and Statistical Tests

Global test accuracy measures population-level predictive utility. Mean and median held-out client accuracy describe the center of the client distribution; p10 and worst-client accuracy describe its lower tail; standard deviation and Jain's index describe dispersion. For each experimental cell, seed-level results should be summarized using mean, sample standard deviation, and a 95% bootstrap confidence interval.

Pairwise algorithm comparisons are performed on matched seed-level differences. When a family of comparisons contains multiple hypotheses, a Holm correction controls family-wise error. Effect sizes and confidence intervals accompany p-values. Communication rounds are repeated measurements within an experiment and are not treated as independent samples for significance testing.

D. Hyperparameter Discipline

Algorithm-specific hyperparameters are selected using a predetermined validation procedure and are not tuned on the final test partition. The minimum sensitivity set includes the FedProx proximal coefficient, Ditto regularization coefficient, Per-FedAvg adaptation settings, and FedSAM sharpness radius. The number

of selected clients per round and local epoch count are also varied to test whether qualitative conclusions depend on one optimization budget.

VIII. DISCUSSION

The analytical results explain why several common shortcuts can mislead PFL evaluation. First, a Dirichlet parameter is a generator setting, not a complete empirical characterization. Because realized concentration remains random, method comparisons should be paired on the same partition and should report realized label statistics. Second, Jain's index is a normalized dispersion measure. It is useful for comparing equality of client outcomes but cannot establish good predictive performance without an accompanying accuracy metric. Third, pairing algorithms on common seeds and partitions is not merely convenient; it can directly reduce comparison variance when the resulting outcomes are positively correlated.

These observations also clarify what a convincing personalization result should look like. A method that improves mean personalized accuracy but leaves p10 and worst-client accuracy unchanged has demonstrated average personalization, not broad client benefit. Conversely, a method that improves Jain's index while reducing mean and lower-tail accuracy has made outcomes more equal but not more useful. Reporting both axes prevents a single aggregate number from carrying an interpretation it cannot support.

The need for multi-mechanism heterogeneity is equally important. Dirichlet label skew, pathological label restriction, and quantity skew do not represent the same problem. A method designed to correct objective disagreement may behave differently when the main challenge is local data scarcity. Recent PFL methods increasingly combine representation sharing, personalized aggregation, graph structure, or efficient feature extraction [12], [14], [16], [17]; a fair comparison therefore requires heterogeneity mechanisms that expose different failure modes.

The proposed protocol is intended to support this type of evidence without conflating algorithm design with partition randomness. Its emphasis on matched partitions, lower-tail client metrics, and corrected seed-level statistics is applicable beyond the six algorithms considered here and can be extended to newer personalized or fairness-aware methods as they appear.

IX. LIMITATIONS

The framework focuses on performance-distribution fairness across clients. It does not measure demographic group fairness or individual faimess with respect to protected attributes. Those concepts require datasets containing appropriate attributes and separate ethical and statistical analysis.

The selected benchmarks are image-classification tasks. They provide controlled comparison but do not represent the dependencies present in naturally federated language, recommendation, healthcare, or sensor data. Results obtained with the protocol should therefore be interpreted as evidence about the evaluated conditions rather than as a universal ranking of PFL algorithms.

Synthetic partition generators approximate important forms of heterogeneity but cannot reproduce every real client distribution. Reporting realized partition statistics reduces ambiguity without removing this limitation.

Finally, analytical properties of partitioning and fairness metrics do not substitute for measured algorithm outcomes. Claims about the empirical superiority of any personalized method require completion of the matched experiment matrix under the stated protocol.

X. CONCLUSION

This paper presents a client-centric framework for analyzing personalized federated learning under severe statistical heterogeneity. It establishes three practical analytical facts: symmetric Dirichlet concentration changes monotonically with α while realized partitions remain random; Jain's fairness index is exactly an inverse function of relative client-accuracy dispersion and must be interpreted together with absolute utility; and matched seed-and-partition comparisons can reduce variance through positive pairing. Building on these results, the paper defines a multi-dataset protocol that compares shared, drift-corrected, personalized, and sharpness-aware federated methods using global, central, lower-tail, and fairness metrics. The central methodological conclusion is that personalization should be evaluated as a distributional client-outcome problem: improvements in average accuracy are insufficient unless the lower tail and relative disparity are examined under the same realized heterogeneity.

ACKNOWLEDGMENT. GENERATIVE AI DISCLOSURE

OpenAI ChatGPT was used to assist with language drafting, structural organization, and presentation across the manuscript. The author directed the scientific scope, selected the methodology, reviewed the mathematical derivations and literature, verified the references used in the manuscript, and accepts full responsibility for the accuracy, originality, and conclusions of the submitted work.

REFERENCES.

[1] A. Z. Tan, H. Yu, L. Cui, and Q. Yang, "Towards personalized federated learning," IEEE Trans. Neural Netw. Learn. Syst., vol. 34, no. 12, pp. 9587–9603, Dec. 2023, doi: 10.1109/TNNLS.2022.3160699.

[2] Z. Qu, X. Li, R. Duan, Y. Liu, B. Tang, and Z. Lu, "Generalized federated learning via sharpness aware minimization," in Proc. 39th Int. Conf. Mach. Learn. (ICML), vol. 162, 2022, pp. 18250–18280.

[3] X. Ma, J. Zhang, S. Guo, and W. Xu, "Layer-wised model aggregation for personalized federated learning," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2022, pp. 10092–10101, doi: 10.1109/CVPR52688.2022.00985.

[4] J. Zhang, Y. Hua, H. Wang, T. Song, Z. Xue, R. Ma, and H. Guan, "FedALA: Adaptive local aggregation for personalized federated learning," in Proc. AAAI Conf. Artif. Intell., vol. 37, no. 9, pp. 11237–11244, 2023, doi: 10.1609/aaai.v37i9.26330.

[5] K. Matsuda, Y. Sasaki, C. Xiao, and M. Onizuka, "Benchmark for personalized federated learning," IEEE Open J. Comput. Soc., vol. 5, pp. 2–13, 2024, doi: 10.1109/OJCS.2023.3332351.

[6] A. M. Abdelmoniem, C.-Y. Ho, P. Papageorgiou, and M. Canini, "A comprehensive empirical study of heterogeneity in federated learning," IEEE Internet Things J., vol. 10, no. 16, pp. 14071–14083, Aug. 2023, doi: 10.1109/JIOT.2023.3250275.

[7] X. Li, S. Zhao, C. Chen, and Z. Zheng, "Heterogeneity-aware fair federated learning," Inf. Sci., vol. 619, pp. 968–986, Jan. 2023, doi: 10.1016/j.ins.2022.11.031.

[8] T. H. Rafi, F. A. Noor, T. Hussain, and D.-K. Chae, "Fairness and privacy preserving in federated learning: A survey," Inf. Fusion, vol. 105, Art. no. 102198, May 2024, doi: 10.1016/j.inffus.2023.102198.

[9] F. Sabah, Y. Chen, Z. Yang, M. Azam, N. Ahmad, and R. Sarwar, "Model optimization techniques in personalized federated learning: A survey," Expert Syst. Appl., vol. 243, Art. no. 122874, Jun. 2024, doi: 10.1016/j.eswa.2023.122874.

[10] X. Zhao, Q. Cui, W. Li, W. Ni, E. Hossain, Q. Z. Sheng, X. Tao, and P. Zhang, "Convergence-privacy-fairness trade-off in personalized federated learning," IEEE Trans. Mach. Learn. Commun. Netw., vol. 3, pp. 246–262, 2025, doi: 10.1109/TMLCN.2025.3528901.

[11] X. Xing, Q. Zhan, X. Xie, Y. Yang, Q. Wang, and G. Liu, "Flexible sharpness-aware personalized federated learning," in Proc. AAAI Conf. Artif. Intell., vol. 39, no. 20, pp. 21707–21715, 2025, doi: 10.1609/aaai.v39i20.35475.

[12] C. Huang, X. Chen, Y. Zhang, and H. Wang, "FedCoSR: Personalized federated learning with contrastive shareable representations for label heterogeneity in non-IID data," IEEE Trans. Cybern., vol. 55, no. 12, pp. 5841–5854, Dec. 2025, doi: 10.1109/TCYB.2025.3599370.

[13] Z. Fan, S. Hu, J. Yao, G. Niu, Y. Zhang, M. Sugiyama, and Y. Wang, "Locally estimated global perturbations are better than local perturbations for federated sharpness-aware minimization," in Proc. 41st Int. Conf. Mach. Learn. (ICML), vol. 235, 2024, pp. 12858–12881.

[14] X. Wu, J. Niu, X. Liu, G. Zhu, S. Tang, W. Lin, and J. Cao, "The diversity bonus: Learning from dissimilar clients in personalized federated learning," IEEE Trans. Neural Netw. Learn. Syst., vol. 36, no. 10, pp. 18613–18627, Oct. 2025, doi: 10.1109/TNNLS.2025.3585927.

[15] T. Wang, K. Zhang, J. Cai, and Y. Gong, "Analyzing the impact of personalization on fairness in federated learning for healthcare," J. Healthcare Informatics Res., vol. 8, no. 2, pp. 181–205, 2024, doi: 10.1007/s41666-024-00164-7.

[16] K. Khan, "FedGraph-Fair: Federated learning with personalization and fairness via dynamic graphs and distributionally robust optimization," Inf. Sci., vol. 728, Art. no. 122710, Feb. 2026, doi: 10.1016/j.ins.2025.122710.

[17] C. Liu, N. Zhou, and W. Yu, "Personalized federated learning algorithm based on composite feature extractor," IEEE Internet Things J., vol. 13, no. 11, pp. 22947–22958, Jun. 2026, doi: 10.1109/JIOT.2026.3667304.