FedUP: Uncertainty-Aware Personalized Federated Learning via Probabilistic Prototypes

Furui Qi[1,2], **Weishan Zhang**[1,2,*], **Lingzhao Meng**[1,2,*], **Yuru Liu**[1,2], **Zijun Feng**[1,2], **Yuange Liu**[1,2], **Baoyu Zhang**[1,2], **Daobin Luo**[1,2] and **Tao Chen**[1,2]

[1]State Key Laboratory of Chemical Safety, Qingdao Institute of Software, College of Computer Science and Technology, China University of Petroleum (East China), Qingdao 266580, China

[2]Shandong Key Laboratory of Intelligent Oil & Gas Industrial Software, China University of Petroleum (East China), Qingdao 266580, China

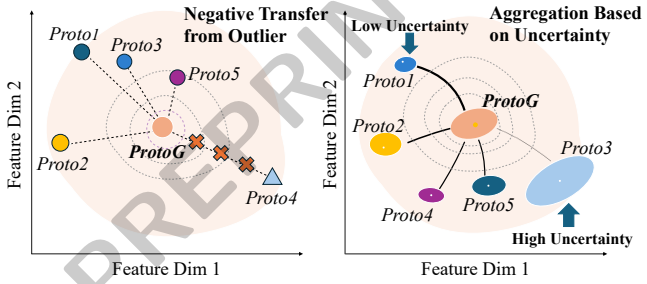
{by2107020320, zhangws}@upc.edu.cn, menglz24@163.com

Abstract

Prototype-based federated learning enables efficient knowledge sharing by exchanging class prototypes rather than full model parameters. However, heterogeneous client data and limited local samples increase prototype estimation variance, making many client prototypes unreliable. Existing methods usually treat prototypes as deterministic point estimates and cannot quantify their reliability, which may contaminate global prototypes and cause negative transfer. To address these challenges, we propose FedUP, an uncertainty-aware personalized federated learning framework that models prototypes as probability distributions. FedUP captures both aleatoric and epistemic uncertainty on a probabilistic simplex to guide local training and global aggregation. It further uses global probabilistic prototypes as class-conditional priors for feature augmentation, alleviating client-side data sparsity. Finally, FedUP aggregates prototype distributions through reliability-informed barycenters on the statistical manifold, suppressing unreliable contributions while preserving geometric structure. Experiments on natural and medical benchmarks demonstrate that FedUP consistently outperforms state-of-the-art methods.

1 Introduction

Personalized Federated Learning (PFL) addresses statistical heterogeneity (Non-IID) by balancing global knowledge sharing with local adaptation, thereby mitigating client drift and performance degradation [Smith *et al.*, 2017; Kairouz and McMahan, 2021]. However, many PFL methods require sharing full model parameters, causing high communication overhead. Prototype-based methods such as FedProto [Tan *et al.*, 2022a] reduce this cost by exchanging class prototypes, which regularize local training through feature-prototype alignment and support heterogeneous architectures.



(a)Traditional: Deterministic Point Aggregation

(b) FedUP: Probabilistic Prototype Aggregation

Figure 1: Comparison of traditional and FedUP prototype aggregation. Traditional point aggregation is easily biased by outliers, causing negative transfer. FedUP performs uncertainty-aware probabilistic aggregation by assigning lower aggregation weights to high-uncertainty prototypes, yielding a more robust global representation.

Despite their efficiency, existing prototype-based methods represent prototypes as deterministic point estimates and implicitly assume uniform reliability across clients. However, under data heterogeneity, prototype quality varies significantly. Clients with scarce or biased data produce unreliable prototypes that poorly represent true class distributions, while data-rich clients generate more accurate ones [Fang *et al.*, 2025; Chen *et al.*, 2025]. Such reliability variation is particularly prominent in medical image analysis, where class imbalance, image noise, and sample-level difficulty often lead to biased or unreliable supervision signals [Meng *et al.*, 2026; Guo *et al.*, 2026]. This heterogeneity in prototype quality, combined with the lack of reliability quantification, leads to two critical problems. **(1) Intra-client: Unreliable prototypes mislead local training.** Existing methods apply fixed-strength alignment for all clients. However, clients with unreliable prototypes should rely more on global guidance, while those with reliable prototypes should preserve local knowledge. Without reliability quantification, existing methods cannot adapt alignment strength accordingly. *How to quantify prototype reliability and adaptively balance global alignment with local personalization* (**Challenge #1**)? **(2) Inter-client: Unreliable prototypes degrade global aggregation.** Exist-

*Corresponding authors: Weishan Zhang and Lingzhao Meng.

ing methods aggregate client prototypes without considering reliability differences. This allows unreliable prototypes from data-scarce clients to dominate the global prototype, causing negative transfer [Wang *et al.*, 2025]. *How to robustly aggregate prototypes with varying reliability while suppressing negative transfer* (**Challenge #2**)?

To address both challenges, we propose FedUP, an uncertainty-aware personalized federated learning framework. Built upon the Dirichlet-based evidential deep learning, FedUP models the categorical prediction of a sample as following a Dirichlet distribution, thus allowing multiple potential predictions for a sample and enabling principled decomposition of predictive uncertainty into epistemic and aleatoric. On the client side, FedUP introduces an adaptive training strategy that leverages decomposed uncertainty for personalization. Epistemic uncertainty is utilized to quantify the model's lack of knowledge for each sample and serves as a dynamic gating mechanism to modulate the alignment intensity between local features and global prototypes. To mitigate the statistical vulnerability caused by the scarcity of local data, FedUP introduces a feature enhancement mechanism. The global probability prototype is regarded as a class-conditional prior, encoding the semantic knowledge of the categories in the federation. While enriching the sparse local representation, it also standardizes the decision boundary of the classifier.

On the server side, as illustrated in Figure 1, unlike traditional methods that aggregate prototypes via naive averaging without considering reliability, FedUP models each client prototype as a Gaussian distribution parameterized by uncertainty-aware statistics, where uncertainty inflates the covariance to reflect prototype reliability. Prototype aggregation is then formulated as computing the barycenter on the statistical manifold, weighted by client reliability derived from both uncertainty types. This geometry-preserving aggregation avoids the variance collapse inherent in naive Euclidean averaging while effectively filtering negative transfer from low-quality prototypes contributed by data-scarce or noisy clients. We summarize our contributions as follows:

- We propose a framework that regards the prototype as a probability distribution, and quantify the reliability of the prototype.
- We design an uncertainty-driven adaptive training mechanism, which dynamically adjusts the global alignment strength and enriches the sparse feature space, balancing personalization and generalization.
- We propose a robust aggregation method based on the optimal transport theory, which can suppress the negative transfer caused by low-quality prototypes while maintaining the prototype geometry.
- Experiments on multiple datasets show that FedUP outperforms state-of-the-art methods under different degrees of data heterogeneity.

2 Related Work

2.1 Heterogeneous and Personalized FL

Statistical heterogeneity (Non-IID) poses a fundamental challenge to federated optimization [Makhija *et al.*, 2024]. Regularization-based methods mitigate client drift by constraining local updates via proximal terms, such as FedProx [Li *et al.*, 2020], or joint optimization objectives like Ditto [Li *et al.*, 2021]. Architecture decoupling approaches support system heterogeneity by separating models into shared and private components. For instance, FedRep [Collins *et al.*, 2021] and FedPer [Arivazhagan *et al.*, 2019] maintain private classifiers, while FedBABU [Oh *et al.*, 2021] freezes the classifier to update only the backbone. Although flexible, these methods are prone to overfitting private modules under data scarcity. Alternatively, Knowledge Distillation (KD) methods like FedMD [Li and Wang, 2019] and FedGen [Zhu *et al.*, 2021] achieve model-agnostic transfer by aligning output distributions or synthesizing pseudo-data. However, KD-based approaches often incur high communication overhead and implicitly assume uniform reliability across client knowledge, failing to account for the varying quality of heterogeneous local data. Although recent generative approaches [Zhang *et al.*, 2024b; Apellániz *et al.*, 2025; Xiao *et al.*, 2025] utilize synthetic data to mitigate data scarcity, they often require training heavy generator models, whereas our approach efficiently leverages probabilistic prototypes as lightweight generators.

2.2 Prototype-based Federated Learning

Prototype-based FL, represented by FedProto [Tan *et al.*, 2022a], has emerged as a communication-efficient paradigm by exchanging class-level feature centroids. Subsequent works have enhanced prototype discriminability through contrastive learning (e.g., in FedProc [Mu *et al.*, 2023] and FedPCL [Tan *et al.*, 2022b]), gradient-based alignment such as FedPAC [Xu *et al.*, 2023], or geometric adjustments like FedTGP [Zhang *et al.*, 2024a] and FedGH [Yi *et al.*, 2023]. These methods successfully reduce communication costs and support heterogeneous architectures by regularizing local features toward global representations [Guo *et al.*, 2024].

Despite this progress, most prototype-based methods still implicitly assume equal reliability across clients. Under data heterogeneity, low-quality prototypes can mislead local alignment and contaminate global aggregation, resulting in negative transfer. Therefore, it remains important to develop a prototype-based FL framework that can quantify prototype uncertainty and perform reliability-aware aggregation.

3 Preliminaries

Consider a central server and K clients $\mathcal{K} = \{1, \dots, K\}$. Client i holds a private dataset $\mathcal{D}_i = \{(\mathbf{x}_j^i, y_j^i)\}_{j=1}^{n_i}$ drawn from a distinct distribution $\mathcal{P}_i(X, Y)$. Due to privacy constraints, only model updates or knowledge summaries can be exchanged. The standard federated optimization objective minimizes the weighted empirical risk:

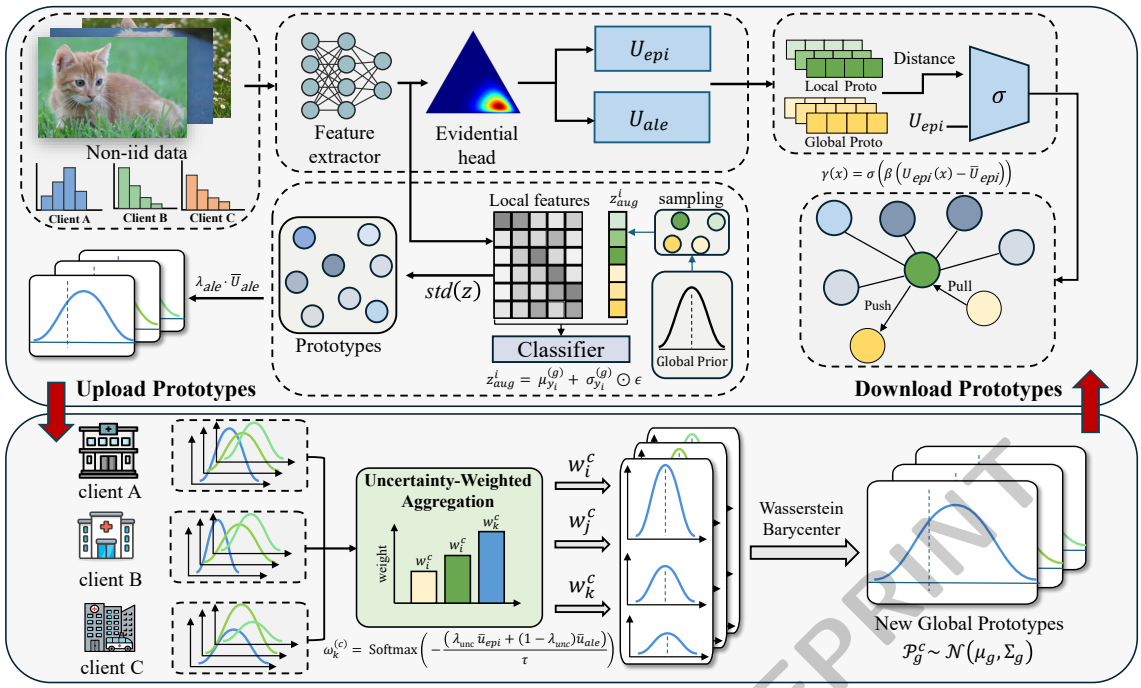


Figure 2: The overall framework of FedUP. It illustrates the client-side evidential uncertainty modeling and the server-side Wasserstein prototype aggregation process.

$$\min_{\{\theta_i\}} \sum_{i=1}^{K} \frac{n_i}{n} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_i}[\ell(f_i(\mathbf{x};\theta_i), y)], \tag{1}$$

where $n = \sum_{i=1}^{K} n_i$, $f_i(\cdot;\theta_i)$ denotes the client model and $\ell(\cdot)$ is the classification loss.

To achieve architecture-agnostic knowledge sharing, existing methods adopt prototype-based representation. Let $h_i(\cdot)$ be the feature extractor of client i. The deterministic prototype for class c is defined as $\mu_i^c = \frac{1}{|\mathcal{D}_i^c|} \sum_{(x,y)\in\mathcal{D}_i^c} h_i(x)$, where $\mathcal{D}_i^c = \{(x,y) \in \mathcal{D}_i \mid y = c\}$. The server aggregates client prototypes via weighted averaging to obtain global prototypes, which are then broadcast to regularize local training.

However, under data heterogeneity, prototypes from data-scarce clients may be highly unreliable yet appear equally credible during aggregation, leading to global knowledge drift. To address this limitation, we extend prototypes from points to probability distributions in Section 4.

4 Methodology

We present **FedUP** to address the two challenges by extending prototypes from deterministic points to probabilistic distributions. The core idea is to quantify prototype reliability via uncertainty. As shown in Figure 2, we employ Evidential Deep Learning to decompose predictive uncertainty into epistemic and aleatoric uncertainty. On the client side, epistemic uncertainty dynamically balances global alignment and local personalization, while global priors enrich sparse feature spaces to address **Challenge #1**. On the server side, prototypes are modeled as Gaussians and fused via uncertainty-weighted Wasserstein barycenters [Agueh and Carlier, 2011], preserving geometric structure while filtering unreliable contributions to address **Challenge #2**.

4.1 Local Training with Uncertainty Modeling

Evidential Uncertainty Quantification

Evidential Deep Learning (EDL) models the categorical prediction $\boldsymbol{\rho}$ as a random variable following a Dirichlet distribution $\text{Dir}(\boldsymbol{\rho}|\boldsymbol{\alpha})$ [Malinin and Gales, 2018; Sensoy *et al.*, 2018]. Unlike conventional networks, EDL learns concentration parameters $\boldsymbol{\alpha}$ that enable explicit uncertainty decomposition through a single forward pass. For a given sample $\mathbf{x}$ and model parameters $\boldsymbol{\theta}$, the probability density function of $\boldsymbol{\rho}$ is defined as:

$$p(\boldsymbol{\rho}|\mathbf{x},\boldsymbol{\theta}) = \begin{cases} \frac{\Gamma(\sum_{c=1}^{C}\alpha_c)}{\prod_{c=1}^{C}\Gamma(\alpha_c)} \prod_{c=1}^{C} \rho_c^{\alpha_c - 1}, & \boldsymbol{\rho} \in \Delta^C \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where $\boldsymbol{\alpha}$ is the concentration parameter vector, $\Gamma(\cdot)$ is the Gamma function, and $\Delta^C = \{\boldsymbol{\rho} \in \mathbb{R}^C \mid \sum_{c=1}^{C} \rho_c = 1, \rho_c > 0\}$ denotes the C-dimensional unit simplex. The model prediction for class c, i.e., the expected categorical probability $P(y = c|\mathbf{x},\boldsymbol{\theta})$, is obtained by marginalizing over $\boldsymbol{\rho}$:

$$P(y = c|\mathbf{x},\boldsymbol{\theta}) = \int p(y = c|\boldsymbol{\rho}) \cdot p(\boldsymbol{\rho}|\mathbf{x},\boldsymbol{\theta})\, d\boldsymbol{\rho} = \frac{\alpha_c}{S}, \tag{3}$$

where $S = \sum_{c=1}^{C} \alpha_c$ is the total Dirichlet strength.

In practice, we append an evidential head after the feature extractor, which outputs non-negative evidence $\mathbf{e}(\mathbf{x}) \in \mathbb{R}_+^C$ via softplus activation. The Dirichlet parameters are then computed as:

$$\boldsymbol{\alpha}(\mathbf{x}) = \mathbf{e}(\mathbf{x}) + \mathbf{1}, \quad S(\mathbf{x}) = \sum_{c=1}^{C} \alpha_c(\mathbf{x}). \tag{4}$$

These Dirichlet parameters enable us to explicitly quantify the predictive uncertainty through a single forward pass

[Chen and Zhang, 2024]. The total predictive uncertainty can be decomposed into two components: aleatoric uncertainty and epistemic uncertainty. Aleatoric uncertainty U_{ale} measures the inherent randomness in the data, which cannot be reduced by collecting more samples, reflecting factors such as class overlap or inherent noise. Epistemic uncertainty U_{epi} quantifies the model's lack of knowledge when facing unfamiliar data, which can be reduced as more training data is accumulated [Chen *et al.*, 2024a; Chen *et al.*, 2024b].

We first compute the total uncertainty as the entropy of the expected categorical distribution:

$$U_{\text{total}}(\mathbf{x}) = -\sum_{c=1}^{C} \frac{\alpha_c}{S} \log \frac{\alpha_c}{S}. \tag{5}$$

The aleatoric uncertainty captures the expected entropy under the Dirichlet distribution:

$$U_{\text{ale}}(\mathbf{x}) = -\sum_{c=1}^{C} \frac{\alpha_c}{S} \left(\psi(\alpha_c + 1) - \psi(S + 1)\right), \tag{6}$$

where $\psi(\cdot)$ is the digamma function. Finally, the epistemic uncertainty is obtained by subtracting aleatoric uncertainty from total uncertainty:

$$U_{\text{epi}}(\mathbf{x}) = U_{\text{total}}(\mathbf{x}) - U_{\text{ale}}(\mathbf{x}). \tag{7}$$

We train with the following EDL loss:

$$\mathcal{L}_{\text{EDL}} = \mathbb{E}_{\mathcal{D}_i}\left[\mathcal{L}_{\text{CE}}^{\text{Dir}} + \lambda_{\text{KL}} \text{KL}\big(\text{Dir}(\tilde{\boldsymbol{\alpha}}) \| \text{Dir}(\mathbf{1})\big)\right]. \tag{8}$$

where $\mathcal{L}_{\text{CE}}^{\text{Dir}}$ is the expected cross-entropy under the Dirichlet distribution, and the KL term regularizes the evidence of incorrect classes with $\tilde{\boldsymbol{\alpha}} = \boldsymbol{\alpha} - \mathbf{e}_y + \mathbf{1}$.

Probabilistic Prototype Construction

We construct probabilistic prototypes by leveraging aleatoric uncertainty to down-weight noisy samples and to encode class-wise reliability. For client i and class c, let $\mathbf{z}_j = f_\theta(\mathbf{x}_j) \in \mathbb{R}^d$ denote the feature embedding for sample $j \in \mathcal{D}_i^c$. We assign a reliability weight $w_j = \exp(-U_{\text{ale}}^{(j)})$ and compute the weighted prototype mean:

$$\boldsymbol{\mu}_i^c = \frac{\sum_{j \in \mathcal{D}_i^c} w_j \mathbf{z}_j}{\sum_{j \in \mathcal{D}_i^c} w_j}. \tag{9}$$

To model uncertainty in the prototype spread, we first compute the per-dimension empirical standard deviation $\boldsymbol{\sigma}_{\text{feat}}^c = \text{Std}(\{\mathbf{z}_j\}_{j \in \mathcal{D}_i^c})$ and then inflate it by class average aleatoric uncertainty:

$$\boldsymbol{\sigma}_i^c = \boldsymbol{\sigma}_{\text{feat}}^c + \lambda_{\text{ale}} \cdot \bar{U}_{\text{ale}}^{(i,c)}, \quad \bar{U}_{\text{ale}}^{(i,c)} = \frac{1}{|\mathcal{D}_i^c|} \sum_{j \in \mathcal{D}_i^c} U_{\text{ale}}^{(j)}. \tag{10}$$

Finally, we represent the class prototype as a diagonal Gaussian distribution:

$$\mathcal{P}_i^c = \mathcal{N}\big(\boldsymbol{\mu}_i^c, \text{diag}((\boldsymbol{\sigma}_i^c)^2)\big), \tag{11}$$

where larger $\bar{U}_{\text{ale}}^{(i,c)}$ yields a wider prototype distribution, suppressing unreliable or noisy classes in subsequent aggregation and alignment.

Dynamic Personalization via Epistemic Uncertainty

A fundamental tension in personalized FL is balancing global knowledge absorption against local pattern preservation. We resolve this by using epistemic uncertainty as an *adaptive gate*: samples where the model lacks knowledge should leverage global guidance, while confident predictions should retain local specificity. For each sample $(\mathbf{x}, y)$, we define a dynamic alignment coefficient:

$$\gamma(\mathbf{x}) = \sigma\left(\beta \cdot \big(U_{\text{epi}}(\mathbf{x}) - \bar{U}_{\text{epi}}\big)\right), \tag{12}$$

where $\sigma(\cdot)$ is the sigmoid function, $\beta > 0$ controls sensitivity, and $\bar{U}_{\text{epi}}$ is a running average providing normalization. This design ensures: high U_{epi} leads to $\gamma \to 1$ (strong alignment with global prototype), while low U_{epi} leads to $\gamma \to 0$ (preserve local representation).

Given the adaptive coefficient, we define the alignment loss as follows. Let $\mathbf{z} = f_\theta(\mathbf{x}) \in \mathbb{R}^d$ denote the feature embedding and $\mathcal{P}_g^c = \mathcal{N}(\boldsymbol{\mu}_g^c, \boldsymbol{\Sigma}_g^c)$ the global probabilistic prototype for class c. To respect the distributional structure, we define the alignment loss using the Mahalanobis distance:

$$\mathcal{L}_{\text{align}}(\mathbf{z}, \mathcal{P}_g^y) = \frac{1}{2}\left(\boldsymbol{\delta}^\top (\boldsymbol{\Sigma}_g^y)^{-1} \boldsymbol{\delta} + \log |\boldsymbol{\Sigma}_g^y|\right), \tag{13}$$

where $\boldsymbol{\delta} = \mathbf{z} - \boldsymbol{\mu}_g^y$ denotes the residual vector. Under our diagonal covariance assumption, the inverse $(\boldsymbol{\Sigma}_g^y)^{-1} = \text{diag}(1/(\sigma_{g,1}^y)^2, \ldots, 1/(\sigma_{g,d}^y)^2)$ can be efficiently computed. Unlike naive Euclidean distance $\|\mathbf{z} - \boldsymbol{\mu}_g^y\|^2$, Mahalanobis distance automatically down-weights high global variance dimensions while emphasizing stable features.

Global Prior-Guided Feature Augmentation

The adaptive alignment mechanism is effective when local data are sufficient. In data-scarce clients, some classes contain too few samples to form reliable prototypes. Therefore, we use global probabilistic prototypes as generative priors. For each class c with insufficient local samples ($|\mathcal{D}_i^c| < n_{\min}$), we draw virtual features from the global Gaussian:

$$\tilde{\mathbf{z}}_c = \boldsymbol{\mu}_g^c + (\boldsymbol{\Sigma}_g^c)^{1/2} \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{14}$$

We feed $\tilde{\mathbf{z}}_c$ into the local classifier head ϕ_i and optimize an auxiliary cross-entropy loss

$$\mathcal{L}_{\text{aug}} = \ell_{\text{CE}}(\phi_i(\tilde{\mathbf{z}}_c), c). \tag{15}$$

This term encourages stable decision boundaries for sparse classes and promotes alignment with the global semantic.

The client objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{EDL}} + \lambda_{\text{align}} \cdot \mathbb{E}_{(\mathbf{x},y)}[\gamma(\mathbf{x}) \cdot \mathcal{L}_{\text{align}}] + \mathcal{L}_{\text{aug}}. \tag{16}$$

Here, $\mathcal{L}_{\text{EDL}}$ trains the evidential head, $\gamma(\mathbf{x})$ scales alignment by epistemic uncertainty, and $\mathcal{L}_{\text{aug}}$ provides prior-guided augmentation. The hyperparameter λ_{align} controls the strength of global prototype alignment.

4.2 Server-Side Prototype Aggregation

After local training, the server collects probabilistic prototypes from participating clients and aggregates them into global prototypes. The key challenge is to handle prototypes with varying reliability while preserving their distributional

Method	CIFAR-10			PathMNIST			Kvasir-Capsule		
	$\alpha = 0.1$	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 0.1$	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 0.1$	$\alpha = 0.3$	$\alpha = 0.5$
FedAvg	56.67±1.24	57.43±3.05	63.09±0.88	43.23±2.15	44.58±1.87	54.30±1.34	53.84±2.41	61.25±1.95	65.22±1.12
FedProx	57.12±1.31	58.21±0.98	64.05±0.76	44.15±2.02	45.32±1.76	55.12±1.29	54.91±2.22	62.10±1.84	66.45±1.05
FedRep	65.43±1.56	68.90±1.12	70.12±0.65	50.12±1.98	55.34±1.45	60.12±1.02	66.23±1.78	70.12±1.33	75.34±0.89
Ditto	86.14±0.95	70.38±1.18	65.16±0.77	88.21±0.82	71.65±1.54	65.67±1.12	94.07±0.65	85.69±1.21	81.81±0.94
FedProto	82.97±0.88	60.95±1.42	52.67±1.05	80.00±1.15	78.81±0.98	69.19±0.85	93.72±0.54	86.14±1.12	82.66±0.78
FedALA	90.03±0.45	82.25±0.67	76.31±0.55	91.39±0.52	83.88±0.71	74.91±0.88	93.12±0.61	89.30±0.58	80.67±0.72
FedPAC	86.46±0.72	79.38±0.85	73.51±0.63	79.29±1.25	66.30±1.66	64.73±1.41	92.17±0.68	89.57±0.74	81.56±0.95
FedGH	86.27±0.81	68.56±1.35	57.13±0.92	84.46±0.93	76.48±1.18	54.49±1.55	94.42±0.48	90.09±0.62	82.76±0.84
FedTGP	75.95±1.05	73.84±0.92	62.47±0.81	89.39±0.76	78.08±1.12	74.77±0.95	95.41±0.35	88.46±0.78	82.86±0.69
FedWBA	86.72±0.68	71.92±1.04	62.21±0.89	86.20±0.85	73.89±1.22	72.76±0.91	92.43±0.55	89.51±0.82	80.58±0.76
FedUP	**90.12**±0.32	**83.89**±0.28	**77.18**±0.19	**93.29**±0.41	**87.04**±1.35	**85.38**±0.22	**95.94**±0.25	**91.58**±0.31	**88.38**±0.18

Table 1: Performance comparison (Top-1 Accuracy % ± Std) on CIFAR-10, PathMNIST, and Kvasir-Capsule under different Non-IID settings (α). The best results are **bold**.

geometry. We address this through two steps: first, we assess prototype reliability using both epistemic and aleatoric uncertainty to compute aggregation weights; second, we aggregate prototypes via Wasserstein barycenters to avoid variance collapse inherent in Euclidean averaging.

Reliability Assessment via Uncertainty. Client prototypes vary in reliability depending on local data conditions. Clients with limited samples produce prototypes with high epistemic uncertainty, while clients with noisy or ambiguous data exhibit high aleatoric uncertainty. Traditional sample-count weighting cannot distinguish these cases. We leverage both uncertainty components for reliability assessment. Each client i reports class-wise uncertainties $\bar{U}_{\text{epi}}^{(i,c)}$ and $\bar{U}_{\text{ale}}^{(i,c)}$ alongside its local prototype $\mathcal{P}_i^c$. The aggregation weight is then computed as:

$$\omega_i^c = \frac{\exp\left(-(\lambda_{\text{unc}}\bar{U}_{\text{epi}}^{(i,c)} + (1-\lambda_{\text{unc}})\bar{U}_{\text{ale}}^{(i,c)})/\tau\right)}{\sum_{j\in\mathcal{S}_c}\exp\left(-(\lambda_{\text{unc}}\bar{U}_{\text{epi}}^{(j,c)} + (1-\lambda_{\text{unc}})\bar{U}_{\text{ale}}^{(j,c)})/\tau\right)} \tag{17}$$

where $\mathcal{S}_c$ denotes clients possessing class c, τ is a temperature parameter, and λ_{unc} balances the relative contributions of epistemic and aleatoric uncertainty in reliability estimation.

Wasserstein Barycenter Aggregation. Standard Euclidean averaging of Gaussian distributions suffers from variance collapse, where the aggregated covariance underestimates true class spread. To preserve distributional geometry, we aggregate on the 2-Wasserstein manifold, weighted by our uncertainty metrics. To aggregate the probabilistic prototypes $\{\mathcal{P}_i^c\}_{i\in\mathcal{S}_c}$ into a global prototype $\mathcal{P}_g^c$, we minimize the weighted accumulation of squared 2-Wasserstein distances:

$$\mathcal{P}_g^c = \arg\min_{\mathcal{P}} \sum_{i\in\mathcal{S}_c} \omega_i^c \cdot W_2^2(\mathcal{P}, \mathcal{P}_i^c), \tag{18}$$

where the distance between two Gaussians is defined as:

$W_2^2(\mathcal{P}_1, \mathcal{P}_2) = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2^2 + \text{tr}(\boldsymbol{\Sigma}_1) + \text{tr}(\boldsymbol{\Sigma}_2) - 2\text{tr}((\boldsymbol{\Sigma}_1^{1/2}\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_1^{1/2})^{1/2})$ [Li *et al.*, 2024]. The solution to this barycenter problem yields the global mean and covariance updates. The barycentric mean is simply the weighted average $\boldsymbol{\mu}_g^c = \sum_{i\in\mathcal{S}_c}\omega_i^c\boldsymbol{\mu}_i^c$. The barycentric covariance $\boldsymbol{\Sigma}_g^c$ is the fixed-point of the iteration $\boldsymbol{\Sigma}_{t+1} = \sum_{i\in\mathcal{S}_c}\omega_i^c(\boldsymbol{\Sigma}_t^{1/2}\boldsymbol{\Sigma}_i^c\boldsymbol{\Sigma}_t^{1/2})^{1/2}$, initialized with the arithmetic mean.

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate FedUP on three benchmarks covering one natural image dataset and two medical image datasets: CIFAR-10 (10 classes, 60K natural images), PathMNIST (9 tissue types, 107K medical histology images), and Kvasir-Capsule (14 gastrointestinal findings, 47K capsule endoscopy images).

Non-IID Data Partition. To simulate statistical heterogeneity in federated learning, we partition client data using a Dirichlet distribution $Dir(\alpha)$. For each class c, we sample a proportion vector $\mathbf{q}_c \sim Dir(\alpha)$ and allocate samples to K clients accordingly. The hyperparameter α controls the degree of data heterogeneity: a smaller α implies stronger data skew and heterogeneity. In this experiment, we set $\alpha \in \{0.1, 0.3, 0.5\}$, where $\alpha = 0.1$ represents an extremely skewed heterogeneous scenario.

Compared Methods. To comprehensively evaluate performance, we compare FedUP with 10 state-of-the-art (SOTA) methods categorized into four paradigms. Classical FL includes standard FedAvg [McMahan *et al.*, 2017] and FedProx [Li *et al.*, 2020], which introduces proximal regularization to handle heterogeneity. Personalized FL employs architecture decoupling or multi-task strategies, including FedRep [Collins *et al.*, 2021], Ditto [Li *et al.*, 2021], and FedALA

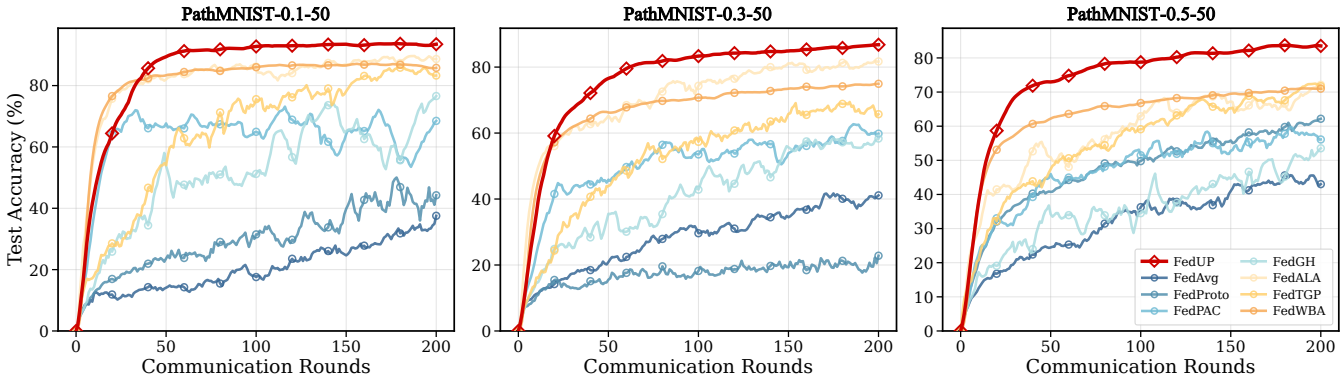


Figure 3: Test accuracy curves over communication rounds on PathMNIST under different Non-IID settings.

Method	HtFE2	HtFE3	HtFE4	HtFE9	Avg.
FedProto	57.78	60.12	36.50	36.77	47.79
FedGH	68.45	70.23	64.89	62.34	66.48
FedTGP	72.11	73.92	75.13	71.55	73.18
FedUP	**77.83**	**75.72**	**77.37**	**73.21**	**76.03**

Table 2: Test accuracy (%) under different levels of model heterogeneity settings on PathMNIST ($\alpha = 0.1$).

Method	$E = 5$	$E = 10$	$E = 20$
FedProto	86.14±0.32	89.80±0.28	89.80±0.35
FedGH	90.09±0.25	89.10±0.31	88.30±0.42
FedALA	89.30±0.27	90.24±0.19	89.20±0.33
FedTGP	90.59±0.21	90.14±0.24	89.95±0.28
FedWBA	89.53±0.30	88.41±0.35	87.92±0.41
FedUP	**90.65±0.25**	**90.29±0.18**	**91.40±0.32**

Table 3: Impact of local training epochs E on PathMNIST ($\alpha = 0.1$).

[Zhang *et al.*, 2023]. Prototype-based FL utilizes feature centroids for knowledge transfer, including FedProto [Tan *et al.*, 2022a], FedPAC [Xu *et al.*, 2023], FedGH [Yi *et al.*, 2023], and FedTGP [Zhang *et al.*, 2024a]. Geometric & Bayesian FL is represented by FedWBA [Wei *et al.*, 2025], which utilizes Wasserstein Barycenter for robust aggregation.

Implementation Details. Regarding model architecture, to ensure fair comparison, we employ ResNet-18 as the backbone for all datasets. For the training configuration, we simulate a federated system with $K = 50$ clients. We set the participation rate to 10% per round and train for a total of $T = 300$ communication rounds. For hyperparameters, local training uses an SGD optimizer with a momentum of 0.9, weight decay of $1e-4$, batch size $B = 10$, and local epochs $E = 5$. The learning rate is initialized at 0.01 and decays using a cosine annealing strategy. Regarding FedUP-specific settings, the evidential Dirichlet prior strength S is initialized to 10. The uncertainty-guided personalization coefficient β is set to 5.0, and the temperature coefficient τ for calculating aggregation weights is set to 0.5.

5.2 Main Results

Model Accuracy. Table 1 reports test accuracy across three datasets under varying non-IID heterogeneity ($\alpha \in \{0.1, 0.3, 0.5\}$). FedUP consistently achieves the best performance across all settings. Notably, on PathMNIST, FedUP significantly outperforms all baselines, achieving **93.29%**, **87.04%**, and **85.38%** under $\alpha = 0.1$, 0.3, and 0.5 respectively. On CIFAR-10, FedUP achieves **90.12%** at $\alpha = 0.1$, comparable to FedALA while outperforming prototype-based methods such as FedProto and FedGH by over 3%. On Kvasir-Capsule, FedUP reaches **95.94%** at $\alpha = 0.1$ and **88.38%** at $\alpha = 0.5$, where the latter exceeds the second-best method by 5.52%, demonstrating stronger robustness under milder heterogeneity. These results validate that uncertainty-guided local adaptation and geometry-preserving global aggregation can effectively address the challenges posed by data heterogeneity in federated learning.

Convergence Performance. Figure 3 illustrates FedUP's superior convergence on PathMNIST. FedUP reaches 80% accuracy within just 20 rounds, approximately 2.5 times as fast as FedProto and FedPAC which require over 50 rounds. More importantly, baseline methods such as FedAvg, FedProto, and FedPAC exhibit significant oscillations under heterogeneous settings, with accuracy fluctuating by up to 10-15% between rounds. In contrast, FedUP maintains a smooth and stable trajectory throughout training. This stability stems from uncertainty-aware adaptation that suppresses unreliable updates and Wasserstein aggregation that prevents global prototype drift.

5.3 Robustness Analysis

Impact of Model Heterogeneity. To evaluate robustness against model heterogeneity, we test FedUP under four scenarios where clients randomly adopt a fixed architecture from a candidate pool: **HtFE2** (4-layer CNN, ResNet18); **HtFE3** (ResNet10/18/34); **HtFE4** (CNN, GoogleNet, MobileNetV2, ResNet18); and **HtFE9** (9 variants from ResNet4 to ResNet152). As shown in Table 2 (PathMNIST, $\alpha = 0.1$), baseline performance deteriorates significantly with increased heterogeneity; notably, FedProto suffers a drastic 21% drop from HtFE2 to HtFE9. In contrast, FedUP main-

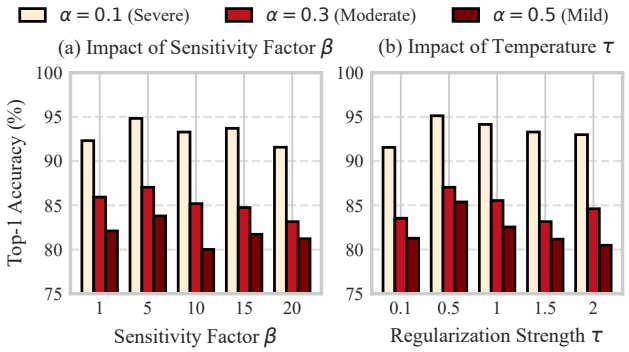


Figure 4: Hyperparameter sensitivity analysis. FedUP consistently achieves optimal performance at $\beta = 5$ and $\tau = 0.5$ across all heterogeneity levels.

Modules			Communication Rounds				
FA	UW	WA	R=50	R=100	R=150	R=200	R=300
✓	-	-	56.45	61.23	65.12	67.45	69.23
-	-	✓	58.34	63.56	67.89	69.82	71.45
-	✓	-	60.12	66.78	70.45	72.89	74.12
✓	-	✓	64.23	70.15	73.45	75.89	77.56
✓	✓	-	69.91	74.47	78.73	80.38	81.05
-	✓	✓	66.92	73.73	76.33	78.52	81.85
✓	✓	✓	**73.26**	**78.87**	**81.84**	**83.64**	**85.38**

Table 4: Ablation study on PathMNIST ($\alpha = 0.5$). We use '-' to denote the absence of a module. FA: Feature Augmentation; UW: Uncertainty Weighting; WA: Wasserstein Aggregation.

tains remarkable stability and achieves 73.21% accuracy in the extreme HtFE9 setting, outperforming the best baseline FedTGP by 1.66%. This validates that our Wasserstein-based probabilistic aggregation aligns heterogeneous feature spaces more effectively than traditional point-based estimation.

Impact of Client Training Epochs. We investigate the robustness of FedUP against client drift by varying local epochs $E \in \{5, 10, 20\}$ on PathMNIST ($\alpha = 0.1$). As shown in Table 3, baselines like FedGH and FedWBA suffer performance degradation as E increases (e.g., FedWBA drops by 1.61%) due to overfitting local heterogeneous data. In contrast, **FedUP** benefits from prolonged local training, achieving peak accuracy of **91.40%** at $E = 20$. This confirms that our uncertainty-guided regularization effectively constrains local updates within a reliable range, preventing severe drift even with extensive local computation.

5.4 Ablation Study

Table 4 presents the ablation study on PathMNIST ($\alpha = 0.5$) with three modules: Feature Augmentation (FA), Uncertainty Weighting (UW), and Wasserstein Aggregation (WA). Among individual components, UW contributes most significantly (74.12% at R=300), validating that uncertainty-guided alignment effectively prevents unreliable updates from corrupting local models, while WA alone (71.45%) outperforms FA alone (69.23%), indicating that geometry-preserving aggregation provides more immediate benefits when prototype quality varies across clients. Combining any two modules yields substantial improvements, where UW+WA achieves 81.85%, demonstrating that client-side uncertainty filtering and server-side geometric aggregation are complementary. The complete FedUP framework achieves 85.38%, outperforming the best two-module variant by 3.53%, confirming that FA mitigates data sparsity, UW enables adaptive personalization, and WA ensures robust knowledge fusion.

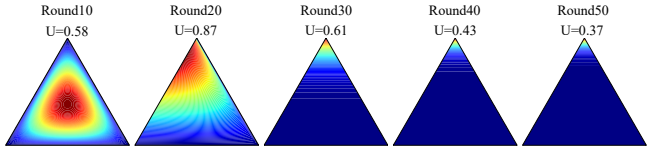


Figure 5: Visualization of the Dirichlet distribution evolution for hard examples on PathMNIST, showing how FedUP gradually reduces uncertainty and improves confidence over 50 training rounds.

5.5 Hyperparameter Sensitivity

Uncertainty-Guided Personalization Coefficient β. β balances local feature preservation and global alignment. As shown in Figure 4(a), performance improves as β increases until $\beta = 5$, beyond which excessive regularization suppresses valid local patterns. This confirms $\beta = 5$ as the optimal trade-off across all heterogeneity levels.

Temperature Coefficient τ. We then analyze the temperature coefficient τ for server-side aggregation weighting. As shown in Figure 4(b), small τ over-concentrates weights on a few clients, reducing diversity, while large τ yields near-uniform weights that fail to filter unreliable prototypes. Setting $\tau = 0.5$ achieves the best balance between reliability filtering and knowledge diversity.

5.6 Visualization of Uncertainty Evolution

Figure 5 visualizes the Dirichlet distribution evolution on the simplex for hard examples on PathMNIST. In early training, the probability mass is dispersed with low evidence strength, indicating high epistemic uncertainty. As training progresses, FedUP strengthens global constraints on uncertain samples and enriches sparse representations through prior-guided augmentation. The distribution gradually concentrates toward the correct class vertex with increasing evidence, suggesting better-calibrated predictions rather than overconfident errors.

6 Conclusion

We propose FedUP, an uncertainty-aware personalized federated learning framework that models class prototypes as probability distributions. By decomposing predictive uncertainty, FedUP adaptively balances local personalization and global alignment, while suppressing unreliable prototype contributions during aggregation. Experiments on natural and medical benchmarks demonstrate the effectiveness and robustness of FedUP under data heterogeneity.

Acknowledgments

This work is supported, in part, by the National Natural Science Foundation of China under Grant Nos. 92567204 and 62072469; the Shandong Provincial Natural Science Foundation Major Basic Research Program under Grant No. ZR2024ZD20; the Shandong Data Open Innovative Application Laboratory.

References

[Agueh and Carlier, 2011] Martial Agueh and Guillaume Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

[Apellániz *et al.*, 2025] Patricia A Apellániz, Juan Parras, and Santiago Zazo. Improving synthetic data generation through federated learning in scarce and heterogeneous data scenarios. *Big Data and Cognitive Computing*, 9(2):18, 2025.

[Arivazhagan *et al.*, 2019] Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*, 2019.

[Chen and Zhang, 2024] Jiayi Chen and Aidong Zhang. On disentanglement of asymmetrical knowledge transfer for modality-task agnostic federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 11311–11319, 2024.

[Chen *et al.*, 2024a] Jiayi Chen, Benteng Ma, Hengfei Cui, and Yong Xia. Fedevi: Improving federated medical image segmentation via evidential weight aggregation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 361–372. Springer, 2024.

[Chen *et al.*, 2024b] Jiayi Chen, Benteng Ma, Hengfei Cui, and Yong Xia. Think twice before selection: Federated evidential active learning for medical image analysis with domain shifts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11439–11449, 2024.

[Chen *et al.*, 2025] Chunlu Chen, Ji Liu, Haowen Tan, Xingjian Li, Kevin I-Kai Wang, Peng Li, Kouichi Sakurai, and Dejing Dou. Trustworthy federated learning: privacy, security, and beyond. *Knowledge and Information Systems*, 67(3):2321–2356, 2025.

[Collins *et al.*, 2021] Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *International conference on machine learning*, pages 2089–2099. PMLR, 2021.

[Fang *et al.*, 2025] Jingdian Fang, Jun Ke, Jiahui Chen, Huiwu Huang, and Shuicai Huang. Fedhcl: Hierarchical cluster fine-grained prototype contrastive learning for multi-feature distributions and model heterogeneity in fl. *IEEE Internet of Things Journal*, 2025.

[Guo *et al.*, 2024] Shunxin Guo, Hongsong Wang, and Xin Geng. Dynamic heterogeneous federated learning with multi-level prototypes. *Pattern Recognition*, 153:110542, 2024.

[Guo *et al.*, 2026] Jie Guo, Lingzhao Meng, Jia Han, Ying Guo, Fengxiang Li, Yipeng Ning, Lishan Qiao, Nianying Sun, Guang Zhang, Xiaoming Xi, et al. Dual-level self-adaptive threshold learning for semi-supervised cnv classification. *Pattern Recognition*, page 113757, 2026.

[Kairouz and McMahan, 2021] Peter Kairouz and H Brendan McMahan. Advances and open problems in federated learning. *Foundations and trends in machine learning*, 14(1-2):1–210, 2021.

[Li and Wang, 2019] Daliang Li and Junpu Wang. Fedmd: Heterogenous federated learning via model distillation. *arXiv preprint arXiv:1910.03581*, 2019.

[Li *et al.*, 2020] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

[Li *et al.*, 2021] Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *International conference on machine learning*, pages 6357–6368. PMLR, 2021.

[Li *et al.*, 2024] Hongxia Li, Wei Huang, Jingya Wang, and Ye Shi. Global and local prompts cooperation via optimal transport for federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12151–12161, 2024.

[Makhija *et al.*, 2024] Disha Makhija, Joydeep Ghosh, and Nhat Ho. A bayesian approach for personalized federated learning in heterogeneous settings. *Advances in Neural Information Processing Systems*, 37:102428–102455, 2024.

[Malinin and Gales, 2018] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 31, 2018.

[McMahan *et al.*, 2017] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[Meng *et al.*, 2026] Lingzhao Meng, Xiaoming Xi, Jia Han, Lishan Qiao, Yilong Yin, and Xinjian Chen. Difficulty-aware pseudo-label correction network for fine-grained classification of choroidal neovascularization in oct images. *IEEE Transactions on Multimedia*, 2026.

[Mu *et al.*, 2023] Xutong Mu, Yulong Shen, Ke Cheng, Xueli Geng, Jiaxuan Fu, Tao Zhang, and Zhiwei Zhang. Fedproc: Prototypical contrastive federated learning on non-iid data. *Future Generation Computer Systems*, 143:93–104, 2023.

[Oh *et al.*, 2021] Jaehoon Oh, Sangmook Kim, and Se-Young Yun. Fedbabu: Towards enhanced representation for federated image classification. *arXiv preprint arXiv:2106.06042*, 2021.

[Sensoy *et al.*, 2018] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.

[Smith *et al.*, 2017] Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S Talwalkar. Federated multi-task learning. *Advances in neural information processing systems*, 30, 2017.

[Tan *et al.*, 2022a] Yue Tan, Guodong Long, Lu Liu, Tianyi Zhou, Qinghua Lu, Jing Jiang, and Chengqi Zhang. Fedproto: Federated prototype learning across heterogeneous clients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 8432–8440, 2022.

[Tan *et al.*, 2022b] Yue Tan, Guodong Long, Jie Ma, Lu Liu, Tianyi Zhou, and Jing Jiang. Federated learning from pre-trained models: A contrastive learning approach. *Advances in neural information processing systems*, 35:19332–19344, 2022.

[Wang *et al.*, 2025] Guowei Wang, Changxing Ding, Wentao Tan, and Mingkui Tan. Decoupled prototype learning for reliable test-time adaptation. *IEEE Transactions on Multimedia*, 2025.

[Wei *et al.*, 2025] Ting Wei, Biao Mei, Junliang Lyu, Renquan Zhang, Feng Zhou, and Yifan Sun. Personalized bayesian federated learning with wasserstein barycenter aggregation. *arXiv preprint arXiv:2505.14161*, 2025.

[Xiao *et al.*, 2025] Yunpeng Xiao, Dengke Zhao, Xufeng Li, Tun Li, Rong Wang, and Guoyin Wang. A federated learning-based data augmentation method for privacy preservation under heterogeneous data. *IEEE Transactions on Mobile Computing*, 2025.

[Xu *et al.*, 2023] Jian Xu, Xinyi Tong, and Shao-Lun Huang. Personalized federated learning with feature alignment and classifier collaboration. *arXiv preprint arXiv:2306.11867*, 2023.

[Yi *et al.*, 2023] Liping Yi, Gang Wang, Xiaoguang Liu, Zhuan Shi, and Han Yu. Fedgh: Heterogeneous federated learning with generalized global header. In *Proceedings of the 31st ACM international conference on multimedia*, pages 8686–8696, 2023.

[Zhang *et al.*, 2023] Jianqing Zhang, Yang Hua, Hao Wang, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. Fedala: Adaptive local aggregation for personalized federated learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 11237–11244, 2023.

[Zhang *et al.*, 2024a] Jianqing Zhang, Yang Liu, Yang Hua, and Jian Cao. Fedtgp: Trainable global prototypes with adaptive-margin-enhanced contrastive learning for data and model heterogeneity in federated learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 16768–16776, 2024.

[Zhang *et al.*, 2024b] Jianqing Zhang, Yang Liu, Yang Hua, and Jian Cao. An upload-efficient scheme for transferring knowledge from a server-side pre-trained generator to clients in heterogeneous federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[Zhu *et al.*, 2021] Zhuangdi Zhu, Junyuan Hong, and Jiayu Zhou. Data-free knowledge distillation for heterogeneous federated learning. In *International conference on machine learning*, pages 12878–12889. PMLR, 2021.