Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-26)
Special Track on AI and Health

One-Shot Federated Class-Incremental Learning for Medical Imaging via Variational Feature Transfer

Pedro H. Barros[1,*], **Omid Orang**[1,*], **Giulia Zanon de Castro**[1], **Heitor S. Ramos**[1] and **Frederico Gadelha Guimarães**[1]

[1]Future Lab, Department of Computer Science, Federal University of Minas Gerais, Belo Horizonte, Brazil
{pedro.barros, omid.orang, ramosh}@dcc.ufmg.br, {giuliaz, fredericoguimaraes}@ufmg.br

Abstract

Federated learning (FL) enables privacy-preserving collaboration for medical image analysis across decentralized institutions but faces major challenges from non-IID data distributions, high communication overhead in multi-round protocols, and catastrophic forgetting when models must adapt to sequentially arriving tasks. These issues are particularly critical in healthcare, where repeated client participation is often infeasible. We address this setting by proposing a novel class-incremental continual learning (CL) model for a one-shot FL paradigm, in which each task introduces new classes, clients observe heterogeneous and evolving class distributions, and communication with the server occurs only once. Clients estimate class-conditional feature distributions via variational inference (VI) from private data and transmit compact statistics to the server, which synthesizes features to train a global classifier in a single communication round. The server aggregates these distributional summaries, synthesizes feature embeddings, and learns a global classifier without revisiting real past data or contacting clients again. Our major novelty is continual adaptation at the distribution level via synthetic replay from stored class mixtures, complemented by lightweight distillation. This approach substantially mitigates catastrophic forgetting while consistently enhancing recognition of newly introduced classes. Extensive experiments on multiple medical imaging benchmarks demonstrate that our method outperforms both state-of-the-art one-shot FL and class-incremental CL approaches. Compared with existing CL models, it achieves 90–97% average accuracy with near-zero forgetting ($\leq 0.10\%$) across different tasks and evaluation settings.

1 Introduction

Deep learning has achieved remarkable success in medical image analysis, enabling accurate diagnosis and decision support across modalities such as X-ray, CT, MRI, dermoscopy, and histopathology [Shen *et al.*, 2017; Chan *et al.*, 2020]. However, training medical imaging models remains challenging due to privacy regulations, limited data sharing across institutions, and data heterogeneity arising from differences in acquisition protocols, devices, patient populations, and disease prevalence [Kaissis *et al.*, 2020].

FL addresses these challenges by enabling collaborative training without centralizing raw data [McMahan *et al.*, 2016]. In FL, clients such as hospitals or medical centers train local models on private data and share only model updates with a central server, thereby preserving data privacy and regulatory compliance. Most existing FL methods rely on multi-round communication protocols, such as FedAvg and FedProx [McMahan *et al.*, 2016; Li *et al.*, 2020], which require repeated client participation and synchronized updates. However, in real-world healthcare settings, repeated communication is often impractical due to strict security policies, restricted network connectivity, limited access windows governed by ethical approvals, and increased exposure to privacy and security risks [Kaissis *et al.*, 2020; Eichelberg *et al.*, 2020].

These constraints motivate one-shot FL, where collaborative model training is completed using a single communication round [Guha *et al.*, 2019]. However, without iterative correction, it must contend with severe data heterogeneity, class imbalance, and domain shifts within a single exchange [Gong *et al.*, 2021; Jin *et al.*, 2022; Zhang *et al.*, 2022]. These methods are often fragile in medical imaging settings, where feature distributions are complex, data are scarce, and small distributional shifts (particularly in rare or ambiguous cases) can lead to abrupt label changes [Kang *et al.*, 2025].

Addressing such evolving settings requires CL, which aims to incrementally incorporate new tasks while mitigating catastrophic forgetting [Parisi *et al.*, 2019; Kirkpatrick *et al.*, 2017]. Although Federated Continual Learning (FCL) has been studied in multi-round and data-accessible settings [Hamedi *et al.*, 2025; Ma *et al.*, 2022; Yang *et al.*, 2024], its integration with one-shot FL remains largely unexplored. Existing FCL methods typically assume repeated client participation, access to real data or gradients, or multiple communication rounds assumptions that are incompatible with one-shot FL and particularly restrictive in privacy-sensitive medical environments.

In this work, we bridge this gap by proposing *VCL-FPFT* (Variational Continual Learning for Federated Parametric Feature Transfer), a novel one-shot FCL framework for medical

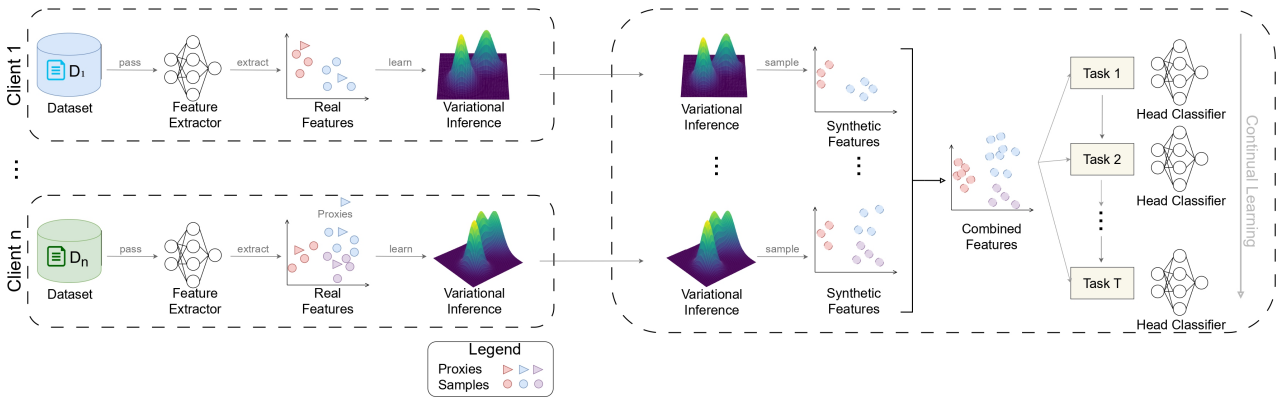

Figure 1: Overview of the proposed one-shot FCL framework. Clients share variational posterior parameters (μ, π, Σ) instead of raw data, enabling privacy-preserving synthetic feature aggregation at the server, followed by class-incremental CL on the head classifier.

imaging. We consider a class-incremental FL setting in which each client participates in only one communication round, historical data are unavailable, and the label space can evolve over time. Our approach operates entirely at the level of feature distributions: clients transmit uncertainty-aware summaries of class-conditional feature statistics, and the server performs distribution-driven continual adaptation using synthetic feature replay and lightweight knowledge distillation. This design enables incremental model adaptation without additional client participation, access to past data, or repeated communication.

Extensive experiments on multiple medical imaging benchmarks under IID and non-IID partitions show that the proposed method achieves robust performance in both one-shot and continual settings, outperforming existing one-shot FL and FCL baselines while maintaining strict privacy and communication constraints.

Our main contributions are

- To the best of our knowledge, we formulate the first one-shot FCL problem for medical imaging, combining single-round communication with class-incremental task evolution.
- Propose a distribution-level continual adaptation mechanism based on uncertainty-aware synthetic feature replay and knowledge distillation, mitigating catastrophic forgetting without access to real data or additional communication rounds.

2 Related Work

This section reviews federated learning, continual learning, and their intersection, focusing on communication efficiency, data heterogeneity, and privacy constraints. We highlight the limitations of existing approaches and motivate the need for one-shot federated continual learning.

Federated Learning. In a standard FL setup, each client performs local optimization on its private dataset and periodically transmits model updates to a coordinating server, which aggregates them to form a global model [Jin *et al.*, 2022]. Most classical FL algorithms, including FedAvg [McMahan *et al.*, 2016] and FedProx [Li *et al.*, 2020], rely on iterative, multi-round communication to mitigate client heterogeneity and improve convergence. While effective, repeated communication incurs significant bandwidth and computational overhead, increases attack surfaces, and may violate deployment constraints in large-scale or privacy-critical environments [Zhang *et al.*, 2021a; Banabilah *et al.*, 2022].

To address these challenges, one-shot FL has been proposed as a communication-efficient alternative, where each client participates in exactly one communication round [Guha *et al.*, 2019; Wang *et al.*, 2025]. Existing one-shot FL methods can be broadly categorized into two paradigms. (i) *Knowledge distillation–based approaches* aggregate client knowledge via logits or soft predictions, often relying on public or proxy datasets [Gong *et al.*, 2021]. (ii) *Parameter fusion or neuron-matching methods* directly align or merge client models without data access [Jin *et al.*, 2022].

Despite their efficiency, existing one-shot FL approaches are primarily designed for static learning problems and remain sensitive to strong statistical heterogeneity. In particular, distillation-based methods typically assume the availability of auxiliary data, while parameter-fusion methods struggle under label skew and task mismatch. Moreover, these methods do not address scenarios where client data evolve sequentially over time, limiting their applicability in non-stationary and continual learning settings.

Continual Learning. CL addresses the challenge of learning from data streams that arrive sequentially over time, where naive training leads to catastrophic forgetting of previously acquired knowledge [Parisi *et al.*, 2019; Kemker *et al.*, 2018]. CL methods aim to enable models to incrementally acquire new knowledge while retaining performance on earlier tasks.

Existing CL approaches can be broadly categorized by their mechanisms for preserving past knowledge. *Regularization-based methods*, such as Elastic Weight Consolidation (EWC) [Kirkpatrick *et al.*, 2017], penalize changes to parameters deemed important for previous tasks. *Parameter isolation methods* allocate task-specific sub-networks or parameters to avoid interference, as in PackNet [Mallya and Lazebnik, 2017] and PathNet [Fernando *et al.*, 2017]. *Dynamic architecture approaches*, including Progressive Neural Networks [Rusu *et al.*, 2022], expand model capacity as new tasks arrive. *Replay-based methods* rehearse past knowledge

Dataset	α	N. Tasks	Overall Acc.	Task 0	Task 1	Task 2	Task 3	Avg. Acc.
TissueMNIST	0.3	4	56.25	82.68	81.28	73.81	73.33	77.78
TissueMNIST	0.6	4	56.55	81.36	81.61	75.01	75.23	78.30
TissueMNIST	IID	4	59.92	83.82	84.14	78.98	78.55	81.37
PathMNIST	0.3	3	90.47	99.25	99.69	91.32	–	96.75
PathMNIST	0.6	3	91.89	99.17	99.91	90.98	–	96.69
PathMNIST	IID	3	91.81	99.21	100.00	92.65	–	97.29
DermaMNIST	0.3	4	61.15	71.60	90.12	80.18	100.00	85.48
DermaMNIST	0.6	4	58.40	69.82	90.95	76.98	100.00	84.44
DermaMNIST	IID	4	64.44	71.01	88.48	82.03	100.00	85.38
BloodMNIST	0.3	4	85.44	98.27	95.28	95.45	99.65	97.16
BloodMNIST	0.6	4	86.79	97.81	95.17	95.45	99.91	97.09
BloodMNIST	IID	4	90.59	98.73	96.18	96.02	99.91	97.71
OCTMNIST	0.3	2	73.30	86.20	84.20	–	–	85.20
OCTMNIST	0.6	2	76.10	89.00	82.40	–	–	85.70
OCTMNIST	IID	2	78.10	91.00	81.20	–	–	86.10

Table 1: Classification accuracy of VCL-FPFT on MedMNIST datasets under Dirichlet-based non-IID (α) and IID partitions.

using stored or generated samples [Rolnick *et al.*, 2019], while *knowledge distillation–based methods* transfer information from previous models without retaining historical data [Heo *et al.*, 2019].

Although effective in centralized settings, many CL techniques rely on assumptions that are incompatible with federated environments, such as persistent access to historical data, task identifiers, or centralized control over model capacity. These limitations motivate adaptations of CL principles to decentralized and privacy-constrained learning systems.

Federated Continual Learning. FCL integrates FL and CL to support collaborative learning over distributed clients whose data evolve sequentially over time [Hamedi *et al.*, 2025; Yang *et al.*, 2024]. In realistic decentralized deployments, clients experience non-stationary and heterogeneous data streams, and raw data sharing is prohibited, making FCL both natural and challenging. FCL inherits challenges from both paradigms. From CL, it must mitigate temporal forgetting, where learning new tasks degrades performance on earlier ones [McCloskey and Cohen, 1989]. From FL, it must address spatial forgetting, which arises when aggregating models trained on heterogeneous client distributions or task sequences [Yang *et al.*, 2024]. Effective FCL methods must therefore enable local continual adaptation while supporting robust global knowledge aggregation.

Prior work has adapted classical CL strategies to federated settings. Regularization-based approaches such as FedCurv [Shoham *et al.*, 2019] constrain important parameters across tasks, while knowledge distillation–based methods, including FedCL [Yao and Sun, 2020], transfer knowledge between successive models without storing historical data. Other approaches incorporate CL-inspired mechanisms into federated aggregation to handle non-IID data and concept drift [Hamedi *et al.*, 2025].

FCL settings can be further classified into synchronous and asynchronous regimes [Yang *et al.*, 2024]. In synchronous FCL, all clients follow a shared task sequence, whereas asynchronous FCL allows clients to encounter tasks in distinct and evolving orders. A particularly challenging instance is federated class-incremental learning, where clients observe disjoint and expanding class sets, and the global model must recognize the union of all observed classes over time.

Despite recent advances, most existing FCL methods rely on multi-round communication, repeated client participation, or access to real or proxy data. As surveyed in [Hamedi *et al.*, 2025], current benchmarks predominantly focus on natural images or sensor data and rarely consider communication-constrained or privacy-critical domains such as medical imaging. Even FCL approaches tailored to medical applications typically assume iterative training protocols [Zhou and Wang, 2025].

These limitations expose a critical gap: the lack of FCL frameworks that simultaneously support one-shot communication, asynchronous continual learning, and strict data privacy. Addressing this gap is essential for deploying FL systems in real-world, privacy-sensitive, and resource-constrained environments.

3 Methodology

In many real-world deployments (e.g., hospitals), a global model is trained once using FL and subsequently deployed, while the underlying label space may continue to evolve over time. After deployment, novel classes can emerge due to previously unseen conditions or changing operational requirements, whereas retraining from scratch or revisiting historical client data is often infeasible because of privacy and communication constraints. To address this problem, we propose a one-shot FCL framework that aggregates client knowledge through probabilistic class-conditional VI feature models, as shown in Figure 1. The global representation is estimated in a single communication round, after which the server performs synthetic feature generation and replay-based CL to incorporate newly introduced classes without accessing raw client data. Throughout the entire pipeline, raw samples remain strictly local to clients, and only VI parameter representations are exchanged, ensuring no raw data sharing.

3.1 Problem Statement

We study class-incremental CL under one-shot federated constraints. A system consists of K clients, where client k owns a private dataset $\mathcal{D}_k = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N_k}$ with labels $y_i \in \{1, \dots, C\}$, and data across clients are non-i.i.d. and potentially class-incomplete. The global label space is revealed sequentially as disjoint tasks $\{\mathcal{C}_t\}_{t=1}^{T}$, where at task t the learner must correctly classify all classes in $\bigcup_{\tau \le t} \mathcal{C}_\tau$. Each client communicates with the server exactly once. The server aggregates client information into a global representation and subsequently performs class-incremental learning without further client interaction. The objective is to learn a global classifier $h_\phi : \mathbb{R}^d \to \{1, \dots, C\}$ that incorporates newly introduced classes while preserving performance on previously observed classes under these constraints.

3.2 Local Representation Learning at Clients

Each client independently trains a feature extractor $f_{\theta_k} : \mathbb{R}^{H \times W \times C_{\text{in}}} \to \mathbb{R}^d$, implemented as an attention-based MobileNet backbone with embedding dimension d[1] [Orang *et al.*, 2025]. Given an input $\mathbf{x}$, the neural network produces an embedding $\mathbf{z} = f_{\theta_k}(\mathbf{x})$, which is subsequently ℓ_2-normalized as

[1]Clients start from a pretrained backbone.

$\hat{\mathbf{z}} = \frac{\mathbf{z}}{\|\mathbf{z}\|_2}$, to stabilize the subsequent probabilistic modeling in feature space

Given a normalized embedding $\hat{\mathbf{z}}_i$ with label y_i, we define the cosine similarity $s(\hat{\mathbf{z}}_i, \boldsymbol{\mu}) = \hat{\mathbf{z}}_i^\top \boldsymbol{\mu} / \|\boldsymbol{\mu}\|_2$. The Proxy-NCA likelihood is defined as

$$p(y_i = c \mid \hat{\mathbf{z}}_i) = \frac{\exp(s(\hat{\mathbf{z}}_i, \boldsymbol{\mu}_{k,c}))}{\sum_{c'=1}^{C} \exp(s(\hat{\mathbf{z}}_i, \boldsymbol{\mu}_{k,c'}))}. \tag{1}$$

Client k refines the embedding space using a Proxy-NCA objective, which clusters embeddings around class-specific proxies. Each class c is associated with a variational proxy $q_{k,c}(\boldsymbol{\mu}) = \mathcal{N}(\boldsymbol{\mu}_{k,c}, \boldsymbol{\Sigma}_{k,c})$, where $\boldsymbol{\mu}_{k,c}$ represents the class center (mean) and $\boldsymbol{\Sigma}_{k,c}$ captures intra-class uncertainty (covariance). To mitigate client drift induced by heterogeneous data distributions, batch normalization layers in the backbone are frozen during training [Li *et al.*, 2021].

The $\boldsymbol{\Sigma}_{k,c} \in \mathbb{R}^{d \times d}$ is a full covariance matrix capturing correlations between feature dimensions. To ensure positive semi-definiteness, the covariance is parameterized via a Cholesky decomposition $\boldsymbol{\Sigma}_k = \boldsymbol{L}_k \boldsymbol{L}_k^\top$. The variational means are initialized as $\boldsymbol{\mu}_k \sim \mathcal{U}(-1/\sqrt{d},\, 1/\sqrt{d})$, while $\boldsymbol{L}_k$ is initialized to yield near-isotropic covariances with small variance, following standard practice in variational Bayesian models, as in [Liu *et al.*, 2019].

To incorporate uncertainty, proxies are treated as latent variables and optimized using a variational Proxy-NCA objective [Kim *et al.*, 2022; Barros *et al.*, 2024],

$$\begin{aligned}\mathcal{L}_{\text{V-PNCA}}(\theta_k, \mu_k, \Sigma_k) = &- \sum_i \mathbb{E}_{\boldsymbol{\mu}_{k,y_i} \sim q_{k,y_i}} \left[\log p(y_i \mid \hat{\mathbf{z}}_i)\right] \\ &+ \sum_c \mathrm{KL}(q_{k,c} \,\|\, p_c)\end{aligned} \tag{2}$$

where p_c is an isotropic Gaussian prior. This objective can be interpreted as learning a posterior over class prototypes, where the expected Proxy-NCA likelihood encourages discriminative alignment, while the KL term regularizes uncertainty under limited local data.

Additionally, the KL regularizer prevents overconfident proxies and stabilizes training when class sample sizes are small. In practice, the expectation is approximated using a MCMC sample (five) per iteration with reparameterization trick [Ganguly and Earp, 2021], generalizing Proxy-NCA by representing each class as a distribution instead of a single point, enabling proxies to capture both the mean and uncertainty of local embeddings for robust aggregation across heterogeneous clients.

3.3 One-Shot Server-Side Aggregation

All subsequent embedding extraction and probabilistic modeling are performed in this shared space, ensuring alignment across clients. For each class c, it receives a set of Gaussian components $\{(\boldsymbol{\mu}_{k,c}, \boldsymbol{\Sigma}_{k,c}, n_{k,c})\}_{k \in \mathcal{K}_c}$ from clients containing class c. Instead of averaging parameters, which would obscure client-specific structure, the server constructs a mixture distribution

$$\begin{aligned} q_c^{\text{global}}(\mathbf{z}) &= \sum_{k \in \mathcal{K}_c} \pi_{k,c}\, \mathcal{N}(\mathbf{z} \mid \boldsymbol{\mu}_{k,c}, \boldsymbol{\Sigma}_{k,c}), \\ \pi_{k,c} &= \frac{n_{k,c}}{\sum_{k' \in \mathcal{K}_c} n_{k',c}}. \end{aligned} \tag{3}$$

This aggregation preserves multimodality arising from client heterogeneity and avoids collapsing distinct local representations into a single prototype. In addition, the server estimates a global class prior $p_c = \beta\, \hat{p}_c + (1 - \beta)\frac{1}{C}$, where $\hat{p}_c$ is the empirical class frequency aggregated across clients and $\beta \in (0, 1)$ is a mixing coefficient. The resulting prior is clipped and renormalized to prevent dominance by frequent classes. Intuitively, the server treats each client as a mixture component, preserving multi-modal class structure induced by heterogeneous data.

3.4 Continual Learning with Synthetic Replay

We consider a class-incremental learning setting in which the global class set is partitioned into a sequence of disjoint tasks $\{\mathcal{C}_t\}_{t=1}^{T}$. At task t, new classes $\mathcal{C}_t$ previously unseen was introduced, expanding the total label space from $\mathcal{C}_{t-1}$ to $\mathcal{C}_{\text{total}} = \mathcal{C}_{t-1} \cup \mathcal{C}_t$, with $C_{\text{total}} = C_{t-1} + C_t$. Throughout CL, the backbone encoder remains frozen, and all adaptation is confined to the classifier head.

For each class c, the server maintains a global variational mixture q_c^{global} obtained via feature-level aggregation across clients. When learning task t, synthetic embeddings for newly introduced classes $c \in \mathcal{C}_t$ are sampled from their corresponding global mixtures. To mitigate catastrophic forgetting of previously learned classes, the server additionally generates replay embeddings for classes $c \in \mathcal{C}_{t-1}$ from their stored mixtures.

Let $\{\mathbf{z}_i^{\text{cur}}, y_i^{\text{cur}}\}_{i=1}^{m_{\text{cur}}}$ denote synthetic embeddings from the current task and $\{\mathbf{z}_j^{\text{rep}}, y_j^{\text{rep}}\}_{j=1}^{m_{\text{rep}}}$ replay embeddings from past tasks. At task t, the server trains a new classifier head $h_{\phi_t} : \mathcal{Z} \to \mathbb{R}^{C_{\text{total}}}$ using a combination of cross-entropy and knowledge distillation losses [Hinton *et al.*, 2014]. Specifically, the training objective is

$$\begin{aligned}\mathcal{L}(\phi) = &\underbrace{\frac{1}{m_{\text{cur}}} \sum_{i=1}^{m_{\text{cur}}} \ell_{\text{CE}}(h_\phi(\mathbf{z}_i^{\text{cur}}), y_i^{\text{cur}})}_{\mathcal{L}_{\text{CE,cur}}(\phi)} + \underbrace{\frac{1}{m_{\text{rep}}} \sum_{j=1}^{m_{\text{rep}}} \ell_{\text{CE}}\big(h_\phi(\mathbf{z}_j^{\text{rep}}), y_j^{\text{rep}}\big)}_{\mathcal{L}_{\text{CE,rep}}(\phi)} \\ &+ \underbrace{\lambda_{\text{KD}}\, T^2 \frac{1}{m_{\text{rep}}} \sum_{j=1}^{m_{\text{rep}}} D_{\text{KL}}\left(\sigma\left(\frac{h_\phi(\mathbf{z}_j^{\text{rep}})}{T}\right) \,\middle\|\, \sigma\left(\frac{h_{\phi_{t-1}}(\mathbf{z}_j^{\text{rep}})}{T}\right)\right)}_{\mathcal{L}_{\text{KD}}(\phi)}.\end{aligned} \tag{4}$$

where $\sigma(.)$ is the softmax function, $\ell_{\text{CE}}(\cdot, y) = -\log(\sigma(\cdot)_y)$ is the cross-entropy loss, $h_{\phi_{t-1}}$ is a frozen copy of the classifier head from the previous task, $\lambda_{\text{KD}} > 0$ controls the distillation strength, and $T > 0$ is a temperature parameter.

The cross-entropy terms encourage discrimination over both current and replayed classes, while the distillation loss preserves the output distribution of the previous classifier on replayed embeddings, thereby stabilizing predictions for old classes. By performing CL entirely in the feature space with synthetic replay, the proposed method avoids storing raw data,

incurs minimal communication overhead, and enables scalable class-incremental adaptation at the server[2].

4 Experiments

To our knowledge, there is no public benchmark that simultaneously satisfies: (i) single-round FL, (ii) post-deployment class-incremental expansion, and (iii) no access to historical client data. We therefore evaluate one-shot FL baselines for the deployment stage and CL baselines for the post-deployment stage under a unified protocol and additionally report their performance. To ensure reproducibility, code is available at: https://github.com/Omid-Orang/one-shot-FL-VCL.

Protocol. We follow a two-stage timeline. During deployment, $K = 5$ clients participate in a single communication round: each client trains a local embedding model and transmits only class-conditional VI parameters to the server, while raw samples remain local. The server aggregates client VI parameters to obtain a global feature model and trains an initial classifier. After deployment, the server performs class-incremental updates using synthetic feature replay generated from the aggregated VI model, without revisiting any historical client data.

Datasets. We evaluate on MedMNIST datasets (BloodMNIST, DermaMNIST, OCTMNIST, PathMNIST, OrganAMNIST, TissueMNIST) [Yang *et al.*, 2023], and additionally consider higher-resolution datasets (224×224), including RSNA [Anouk Stein *et al.*, 2018], Diabetic Retinopathy [Dugas *et al.*, 2015], and ISIC [Codella *et al.*, 2019; Kang *et al.*, 2025].

Task construction and data partitions. We follow a class-incremental protocol where the label space is revealed as $T \in \{2, 3, 4\}$ disjoint tasks. At task t, the model must classify all classes in $\cup_{\tau \le t} \mathcal{C}_\tau$. Classes are evenly split across tasks with a fixed ordering per seed. We evaluate both IID and Dirichlet-based non-IID partitions with $\alpha \in \{0.3, 0.6\}$.

Baselines. For one-shot FL, we compare against FedAvg[3] [McMahan *et al.*, 2016], DAFL [Chen *et al.*, 2019], DENSE [Zhang *et al.*, 2021b], FedISCA, and E-FedISCA [Kang *et al.*, 2025]. FedAvg(1) denotes a strict single-epoch local update before aggregation. For post-deployment learning, we benchmark against representative class-incremental learning methods in [Verma *et al.*, 2023; Derakhshani *et al.*, 2022], including regularization-, expansion-, and replay-based approaches under the same task protocol.

Implementation details. Embedding dimension is fixed to 256. Local models are trained for 50 epochs using AdamW with learning rate 10^{-4} and batch size 128. On the server, 2,000 synthetic features per class are generated using a mixed class prior. The global classifier is trained for 100 epochs with learning rate 5×10^{-4}. During class-incremental updates, replay uses 8,000 synthetic samples per previously observed class, with knowledge distillation weight $\lambda_{KD} = 0.2$ and temperature $T = 2.0$. All experiments are conducted on an NVIDIA RTX 6000 GPU.

Metrics. Let $a_{t,i}$ denote the accuracy on task i after training task t. Final average accuracy is defined as $\mathrm{AvgAcc} = \frac{1}{T}\sum_{i=1}^{T} a_{T,i}$. Average forgetting is computed as $F = \frac{1}{T-1}\sum_{i=1}^{T-1}\left[(\max_{t\in\{1,\dots,T\}} a_{t,i}) - a_{T,i}\right]$, following standard CL settings [Verma *et al.*, 2023; Derakhshani *et al.*, 2022]. We additionally report overall accuracy on the union of all observed classes.

5 Results and Discussion

We evaluate VCL-FPFT in three scenarios: (i) comparison with one-shot FL methods at deployment, (ii) robustness under model heterogeneity, and (iii) post-deployment class-incremental learning compared with representative CL baselines.

Scenario I: Comparison with One-Shot FL Methods. We first compare VCL-FPFT with recent one-shot FL approaches under IID and Dirichlet-based non-IID partitions with $\alpha \in \{0.3, 0.6\}$. Table 1 reports the class-incremental performance of VCL-FPFT across MedMNIST datasets. As expected, stronger heterogeneity (smaller α) generally reduces overall and average accuracies, while IID partitions usually yield the best results. Importantly, VCL-FPFT maintains stable accuracy across tasks, indicating that the post-deployment updates do not collapse performance on earlier classes.

Table 2 compares one-shot FL methods under non-IID partitions. VCL-FPFT achieves the best performance on BloodMNIST and PathMNIST under both $\alpha = 0.6$ and $\alpha = 0.3$, and remains best or second-best on OCTMNIST and TissueMNIST ($\alpha = 0.3$). Under severe heterogeneity ($\alpha = 0.3$), VCL-FPFT yields clear gains over FedAvg(1), DAFL, and DENSE, and improves upon FedISCA and E-FedISCA on several datasets. These results suggest that aggregating class-conditional VI parameters provides an effective global feature model even when client data are non-IID.

We further report IID performance across eight datasets in Table 3. VCL-FPFT achieves top accuracy on BloodMNIST and PathMNIST and remains competitive on other MedMNIST benchmarks. On higher-resolution datasets (RSNA, Diabetic Retinopathy, and ISIC), VCL-FPFT also improves on most baselines, showing that our proposed one-shot aggregation and subsequent synthetic-feature replay generalize beyond small, low-resolution datasets.

Discussion. Across heterogeneous and homogeneous partitions, VCL-FPFT consistently provides strong performance while enabling post-deployment class expansion without client revisit. Compared to one-shot FL baselines, we achieved the best performance in 5 out of 10 cases and the second-best in 3 out of 10, supporting our design choice of probabilistic class-conditional modeling and server-side replay.

Scenario II: Impact of Model Heterogeneity. We next evaluate robustness to model heterogeneity, where clients employ heterogeneous local models while still communicating only

[2]Before final classifier training, a variational Proxy-NCA step is applied to synthetic features of newly introduced classes to estimate their class-conditional variational parameters.

[3]FedAvg is evaluated in two variants: standard FedAvg with 50 communication rounds, and a one-shot version (FedAvg(1)) with a single round of aggregation.

Method	Dirichlet ($\alpha = 0.6$)					Dirichlet ($\alpha = 0.3$)				
	Blood	Derma	Oct	Path	Tissue	Blood	Derma	Oct	Path	Tissue
FedAvg [McMahan *et al.*, 2016]	93.60	72.72	76.50	81.48	55.61	87.49	69.88	73.50	77.52	53.26
FedAvg(1)	18.24	66.88	25.00	5.86	32.07	16.92	10.97	25.00	5.86	32.07
DAFL [Chen *et al.*, 2019]	7.13	66.88	34.40	14.97	39.15	7.13	13.62	25.00	18.64	45.00
DENSE [Zhang *et al.*, 2022]	34.52	67.78	39.40	30.31	9.47	30.78	12.77	25.80	19.87	9.33
FedISCA [Kang *et al.*, 2025]	82.90	**69.83**	68.60	82.92	53.04	46.59	15.91	60.50	79.25	51.00
E-FedISCA [Kang *et al.*, 2025]	83.10	69.68	66.20	78.51	52.87	45.86	16.96	61.60	73.40	48.45
FBPFT-VI [Orang *et al.*, 2025]	82.17	63.39	**79.40**	89.21	55.51	83.51	**62.24**	**76.00**	87.08	**56.46**
VCL-FPFT	**86.79**	58.40	76.10	**91.89**	**56.55**	**85.44**	61.15	73.30	**90.47**	56.25

Table 2: Classification accuracy (%) comparison under Dirichlet-based non-IID partitions ($\alpha = 0.6$ and $\alpha = 0.3$) across five MedMNIST datasets. Best results are shown in bold and second-best results are underlined.

Method	Blood	Derma	Oct	Path	Tissue	RSNA	Diabetic	ISIC
FedAvg [McMahan *et al.*, 2016]	93.51	74.61	75.60	84.54	63.64	88.16	49.04	62.88
FedAvg(1)	13.74	66.88	25.00	5.86	32.07	78.65	35.60	38.05
DAFL [Chen *et al.*, 2019]	7.13	66.43	25.00	7.63	11.55	50.55	22.63	14.51
DENSE [Zhang *et al.*, 2022]	39.37	66.93	33.80	21.89	21.35	55.06	23.51	13.69
FedISCA [Kang *et al.*, 2025]	87.99	70.12	70.20	84.18	**61.90**	85.34	40.08	48.39
E-FedISCA [Kang *et al.*, 2025]	87.31	**71.47**	71.30	79.48	57.96	**85.46**	41.32	51.17
FBPFT-VI [Orang *et al.*, 2025]	88.86	67.38	**81.20**	88.34	57.64	84.49	49.54	**68.99**
VCL-FPFT	**90.59**	64.44	78.10	**91.81**	59.92	84.13	**57.80**	67.27

Table 3: Comparative IID Performance (%) on Eight Datasets. The best results for each dataset are highlighted in bold, and the second-best results are underlined.

Dataset	α	**N. Tasks**	**Overall Acc.**	**Task 0**	**Task 1**	**Task 2**	**Task 3**	**Avg. Acc.**
TissueMNIST	0.3	4	59.63	85.82	82.86	76.22	73.14	79.51
TissueMNIST	0.6	4	60.48	82.80	81.89	80.04	78.35	80.77
TissueMNIST	IID	4	64.22	86.03	85.98	82.25	80.44	83.68
PathMNIST	0.3	3	92.05	99.21	99.51	92.23	–	96.98
PathMNIST	0.6	3	92.12	99.41	99.47	93.07	–	97.32
PathMNIST	IID	3	92.90	99.17	97.79	93.65	–	96.87
DermaMNIST	0.3	4	67.13	68.05	93.42	84.72	100.00	86.55
DermaMNIST	0.6	4	62.34	68.64	90.53	77.75	100.00	84.23
DermaMNIST	IID	4	75.16	74.56	91.77	87.66	100.00	88.50
BloodMNIST	0.3	4	92.96	99.65	96.63	97.72	100.00	98.50
BloodMNIST	0.6	4	93.51	99.54	97.08	98.10	100.00	98.68
BloodMNIST	IID	4	95.24	99.77	98.20	98.67	100.00	99.16
OCTMNIST	0.3	2	76.50	90.20	86.20	–	–	88.20
OCTMNIST	0.6	2	77.80	91.20	80.80	–	–	86.00
OCTMNIST	IID	2	77.00	90.00	82.40	–	–	86.20

Table 4: Class-incremental classification accuracy of VCL-FPFT on MedMNIST datasets under non-IID (α) and IID partitions and model heterogeneity.

class-conditional VI parameters.This setting captures realistic FL environments with device variability and non-uniform model configurations.

For model heterogeneity experiments, we followed the setup used in [Kang *et al.*, 2025], where clients used ResNet34, WRN-16-2, VGG16 (BN), and VGG8 (BN). We replaced ResNet18 with our MobileNet-Attention model for one client to represent architectural diversity better.

Table 4 reports the class-incremental performance of VCL-FPFT under heterogeneity. Across all datasets and partitions, VCL-FPFT preserves stable task-wise accuracy and improves both overall and average accuracy compared to the homogeneous setting in Table 1. Table 5 provides a detailed comparison to competing one-shot FL baselines under the same heterogeneous setting. Under IID partitions, VCL-FPFT achieves the best performance on BloodMNIST, DermaMNIST, and TissueMNIST, and remains highly competitive on OCTMNIST and PathMNIST.

Discussion. These results indicate that VCL-FPFT can effectively leverage heterogeneous client representations through a shared probabilistic feature model. In particular, the VI parameters received from clients enables the server to learn a global class-conditional feature distribution that is robust to client-side architectural heterogeneity.

Scenario III: Comparison with Continual Learning Methods. Finally, we evaluate the post-deployment CL stage by comparing VCL-FPFT with representative CL methods following prior CL protocols. We consider baselines spanning regularization-based, expansion-based, and generative/replay-based approaches, as well as FCL methods where applicable. We further extend evaluation to additional MedMNIST datasets under a four-task setting using baselines from [Derakhshani *et al.*, 2022]. Table 6 reports Avg. Acc. and Forgetting averaged over five seeds. VCL-FPFT achieved best Avg. Acc. and near-zero forgetting across BloodMNIST, PathMNIST, OrganaMNIST, and TissueMNIST. Moreover, the low variance across runs indicates stable behavior under class-incremental updates.

Table 7 reports Avg. Acc. and Forgetting on OCTMNIST and PathMNIST (three-task setting). Regularization-based approaches (e.g., EWC, MAS, LwF) exhibit substantial forgetting, and generative approaches only partially mitigate forgetting at a cost in accuracy. In contrast, VCL-FPFT achieves the highest Avg. Acc. while maintaining near-zero forgetting on both datasets. Table 8 further shows task-wise accuracies after the final task, where VCL-FPFT preserves strong performance on earlier tasks without the abrupt degradation observed in competing methods.

Method	IID					Dirichlet ($\alpha = 0.6$)					Dirichlet ($\alpha = 0.3$)				
	Blood	Derma	Oct	Path	Tissue	Blood	Derma	Oct	Path	Tissue	Blood	Derma	Oct	Path	Tissue
DAFL [Chen *et al.*, 2019]	7.13	65.69	25.00	15.72	35.66	7.13	67.21	37.10	28.15	39.54	7.13	13.47	45.30	29.68	19.54
DENSE [Zhang *et al.*, 2022]	46.86	66.88	44.00	33.08	38.28	23.47	67.93	40.70	28.68	36.70	34.67	13.42	44.00	39.37	38.37
FedISCA [Kang *et al.*, 2025]	87.96	71.17	70.00	83.02	**61.74**	73.43	**69.23**	64.80	82.73	51.95	44.20	16.61	62.00	72.26	43.80
E-FedISCA [Kang *et al.*, 2025]	88.31	71.72	71.00	80.04	58.96	72.76	**69.23**	64.60	80.84	52.04	47.18	16.01	58.90	72.17	43.19
FBPFT-VI [Orang *et al.*, 2025]	94.12	66.58	**81.10**	**93.65**	60.81	89.54	57.66	**79.50**	**92.84**	55.23	89.74	65.74	**78.00**	91.34	56.30
VCL-FPFT	**95.24**	**75.16**	77.00	92.90	**64.22**	**93.51**	62.34	77.80	92.12	**60.48**	**92.96**	**67.13**	76.50	**92.05**	**59.63**

Table 5: Classification performance (%) across five datasets under Dirichlet-based non-IID (α) and IID partitions and model heterogeneity. Best and second-best results are shown in bold and underlined, respectively.

Method	**BloodMNIST**		**PathMNIST**		**OrganaMNIST**		**TissueMNIST**	
	Accuracy ↑	Forgetting ↓	Accuracy ↑	Forgetting ↓	Accuracy ↑	Forgetting ↓	Accuracy ↑	Forgetting ↓
LB	46.59±6.46	68.26±8.13	32.29±6.74	77.54±16.40	41.21±7.39	54.20±23.97	21.63±2.55	90.69±1.77
EWC	47.60±7.30	66.22±14.36	33.34±4.77	76.39±15.53	37.88±7.97	67.92±27.40	21.56±2.51	90.48±1.76
MAS	43.94±6.96	69.43±12.90	34.22±6.17	74.98±15.79	44.99±5.61	55.13±24.43	21.11±2.56	89.38±1.72
LwF	43.68±6.55	66.30±8.70	35.36±5.99	67.37±15.76	41.36±6.36	51.47±24.53	21.49±2.58	90.79±1.22
iCaRL	67.70±8.67	14.52±6.93	58.46±8.79	-0.70±6.41	63.02±7.53	7.75±4.49	32.00±3.01	14.42±10.21
EEIL	42.17±7.05	71.25±12.90	28.42±5.43	79.39±15.79	41.03±11.53	62.47±24.43	21.69±2.43	91.35±1.72
VCL-FPFT	**90.92 ± 0.19**	**0.05 ± 0.11**	**91.55 ± 0.61**	**0.01 ± 0.01**	**92.24 ± 0.56**	**-0.01 ± 0.05**	**60.21 ± 0.27**	**0.09 ± 0.13**

Table 6: Comparison of average accuracy and forgetting for class-incremental learning on MedMNIST datasets between our proposed method and baseline CL methods in [Derakhshani *et al.*, 2022] .

Cat.	**Method**	**OCTMNIST**		**PathMNIST**	
		Avg. Acc.	Forgetting	Avg. Acc.	Forgetting
B	Finetune	33.33	100.00	28.89	99.18
B	Joint	90.76	–	89.28	–
R	LwF	44.80	80.23	25.20	81.33
R	LwM	41.75	41.58	23.02	35.43
R	EWC	31.54	76.83	29.16	95.82
R	SI	43.60	21.27	32.40	28.94
R	MAS	31.73	83.69	29.97	74.54
R	MUC-MAS	39.32	76.87	33.49	52.19
R	GPM	41.16	26.41	39.51	24.96
R	OWM	38.93	83.10	52.42	16.34
R	RWalk	33.33	100.00	27.05	85.13
E	EFT	43.20	38.13	66.82	29.39
G	GR	35.83	66.05	21.95	90.88
G	BIR	62.00	51.31	35.17	88.17
F	**VCL-FPFT**	**97.13**	**0.10**	**97.32**	**0.02**

Table 7: Average accuracy and forgetting on OCTMNIST and PathMNIST compared with the methods in [Verma *et al.*, 2023]. Categories: B = Baseline, R = Regularization, E = Expansion, G = Generative, F = Federated Continual Learning.

Category	**Method**	**OCTMNIST**			**PathMNIST**		
		Task 1	**Task 2**	**Task 3**	**Task 1**	**Task 2**	**Task 3**
B	Finetune	0.00	0.00	100.00	0.00	0.00	86.67
B	Joint	99.47	97.47	75.33	95.99	85.93	85.92
R	LwF	0.53	39.33	94.53	0.63	15.09	59.83
R	LwM	63.13	0.93	61.20	28.97	26.69	13.38
R	EWC	0.01	0.13	94.47	1.01	5.20	81.27
R	SI	63.33	14.53	52.93	49.36	24.84	22.99
R	MAS	0.13	10.37	84.68	2.04	6.07	81.78
R	MUC-MAS	9.28	20.99	87.69	18.81	4.14	77.51
R	GPM	90.66	1.73	31.07	79.77	2.23	36.53
R	OWM	14.40	13.60	88.80	77.72	73.18	6.35
R	RWalk	0.00	0.00	100.00	0.22	0.06	80.88
E	EFT	61.47	38.40	29.73	59.60	73.63	67.24
G	GR	2.57	16.81	88.11	0.26	1.82	63.77
G	BIR	10.53	78.40	97.07	17.33	3.38	84.82
F	**VCL-FPFT**	**91.60**	**100.00**	**100.00**	**99.17**	**100.00**	**92.82**

Table 8: Comparison of task-wise classification accuracy (%) under on OCTMNIST and PathMNIST compared with the methods in [Verma *et al.*, 2023]. Categories: B = Baseline, R = Regularization, E = Expansion, G = Generative, F = Federated Continual Learning.

Discussion. Overall, VCL-FPFT offers a trade-off in the post-deployment setting: synthetic feature replay from the aggregated VI model provides plasticity for new classes while distillation-based constraints preserve earlier decision boundaries. This is particularly important in our setting, where the server cannot revisit historical client data and must rely on a compact representation for replay.

6 Conclusion and Future Directions

We addressed a practical post-deployment FL setting where (i) clients participate in only a single communication round, (ii) raw samples remain strictly local, and (iii) the label space expands over time after deployment. Unlike prior FCL approaches that assume repeated client participation, VCL-FPFT introduces class-conditional probabilistic feature summaries as a unifying abstraction that supports one-shot participation and post-deployment class expansion. Extensive experiments on medical benchmarks under IID and Dirichlet non-IID partitions, including model heterogeneity, demonstrate that VCL-FPFT achieves strong performance while maintaining near-zero forgetting during post-deployment class expansion. These results highlight probabilistic feature aggregation as a practical interface between one-shot FL and CL.

Future work will focus on two directions: First, we aim to provide a theoretical characterization of mixture-based aggregation and synthetic replay, including stability and generalization guaranties under client-level distribution shifts and finite variational capacity, clarifying the trade-off between replay budget, approximation error, and forgetting. Second, we plan to extend the framework beyond closed-world class-incremental settings to handle open-set recognition and concept drift, where variational summaries must support uncertainty-aware novelty detection and adaptive mixture reweighting.

Ethical Statement

There are no ethical issues.

Acknowledgments

This work was supported by the Brazilian agencies CNPq (Grants no. 304856/2025-8, "Collaborative Machine Learning and Privacy Protection" and 308380/2025-8) and CAPES through the Academic Excellence Program (PROEX). It is also partially funded by the UFMG–Fundep R&D&I agreement on Federated Machine Learning for Connected Vehicles, and by Kunumi and Embrapii (Project PDCC-2602.0044).

References

[Anouk Stein *et al.*, 2018] MD Anouk Stein, Carol Wu, Chris Carr, George Shih, Jamie Dulkowski, kalpathy, Leon Chen, Luciano Prevedello, MD Marc Kohli, Mark McDonald, Peter, Phil Culliton, Safwan Halabi MD, and Tian Xia. Rsna pneumonia detection challenge. https://kaggle.com/competitions/rsna-pneumonia-detection-challenge, 2018. Kaggle.

[Banabilah *et al.*, 2022] Syreen Banabilah, Moayad Aloqaily, Eitaa Alsayed, Nida Malik, and Yaser Jararweh. Federated learning review: Fundamentals, enabling technologies, and future applications. *Information processing & management*, 2022.

[Barros *et al.*, 2024] Pedro H Barros, Fabricio Murai, Amir Houmansadr, Alejandro C. Frery, and Heitor Soares Ramos Filho. Variational inference in similarity spaces: A bayesian approach to personalized federated learning. In *NeurIPS 2024 Workshop on Bayesian Decision-making and Uncertainty*, 2024.

[Chan *et al.*, 2020] Heang-Ping Chan, Ravi K Samala, Lubomir M Hadjiiski, and Chuan Zhou. Deep learning in medical image analysis. *Deep learning in medical image analysis: challenges and applications*, pages 3–21, 2020.

[Chen *et al.*, 2019] Hanting Chen, Yunhe Wang, Chang Xu, Zhaohui Yang, Chuanjian Liu, Boxin Shi, Chunjing Xu, Chao Xu, and Qi Tian. Data-free learning of student networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3514–3522, 2019.

[Codella *et al.*, 2019] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic), 2019.

[Derakhshani *et al.*, 2022] Mohammad Mahdi Derakhshani, Ivona Najdenkoska, Tom van Sonsbeek, Xiantong Zhen, Dwarikanath Mahapatra, Marcel Worring, and Cees G. M. Snoek. Lifelonger: A benchmark for continual disease classification, 2022.

[Dugas *et al.*, 2015] Emma Dugas, Jared, Jorge, and Will Cukierski. Diabetic retinopathy detection. https://kaggle.com/competitions/diabetic-retinopathy-detection, 2015. Kaggle.

[Eichelberg *et al.*, 2020] Marco Eichelberg, Klaus Kleber, and Marc Kämmerer. Cybersecurity in pacs and medical imaging: an overview. *Journal of Digital Imaging*, 33(6):1527–1542, 2020.

[Fernando *et al.*, 2017] Chrisantha Fernando, Dylan Banarse, Charles Blundell, Yori Zwols, David Ha, Andrei A Rusu, Alexander Pritzel, and Daan Wierstra. Pathnet: Evolution channels gradient descent in super neural networks. *arXiv preprint arXiv:1701.08734*, 2017.

[Ganguly and Earp, 2021] Ankush Ganguly and Samuel W. F. Earp. An introduction to variational inference. *ArXiv*, abs/2108.13083, 2021.

[Gong *et al.*, 2021] Xuan Gong, Abhishek Sharma, Srikrishna Karanam, Ziyan Wu, Terrence Chen, David Doermann, and Arun Innanje. Ensemble attention distillation for privacy-preserving federated learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15076–15086, 2021.

[Guha *et al.*, 2019] Neel Guha, Ameet Talwalkar, and Virginia Smith. One-shot federated learning, 2019.

[Hamedi *et al.*, 2025] Parisa Hamedi, Roozbeh Razavi-Far, and Ehsan Hallaji. Federated continual learning: Concepts, challenges, and solutions. *Neurocomputing*, 651:130844, 2025.

[Heo *et al.*, 2019] Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, 2019.

[Hinton *et al.*, 2014] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *Proceedings of the International Conference on Neural Information Processing Systems*, 1050:9, 2014.

[Jin *et al.*, 2022] Xisen Jin, Xiang Ren, Daniel Preotiuc-Pietro, and Pengxiang Cheng. Dataless knowledge fusion by merging weights of language models. *ArXiv*, abs/2212.09849, 2022.

[Kaissis *et al.*, 2020] Georgios A Kaissis, Marcus R Makowski, Daniel Rückert, and Rickmer F Braren. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2(6):305–311, 2020.

[Kang *et al.*, 2025] Myeongkyun Kang, Philip Chikontwe, Soopil Kim, Kyong Hwan Jin, Ehsan Adeli, Kilian M. Pohl, and Sang Hyun Park. Efficient one-shot federated learning on medical data using knowledge distillation with image synthesis and client model adaptation. *Medical Image Analysis*, 105:103714, 2025.

[Kemker *et al.*, 2018] Ronald Kemker, Marc McClure, Angelina Abitino, Tyler Hayes, and Christopher Kanan. Measuring catastrophic forgetting in neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[Kim *et al.*, 2022] Minyoung Kim, Ricardo Guerrero, Hai X. Pham, and Vladimir Pavlovic. Variational continual proxy-anchor for deep metric learning. In Gustau Camps-Valls,

Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pages 4552–4573, 2022.

[Kirkpatrick *et al.*, 2017] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

[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] Xiaoxiao Li, Meirui JIANG, Xiaofei Zhang, Michael Kamp, and Qi Dou. FedBN: Federated learning on non-IID features via local batch normalization. In *International Conference on Learning Representations*, 2021.

[Liu *et al.*, 2019] Xuanqing Liu, Yao Li, Chongruo Wu, and Cho-Jui Hsieh. Adv-BNN: Improved adversarial defense through robust bayesian neural network. In *International Conference on Learning Representations*, 2019.

[Ma *et al.*, 2022] Yuhang Ma, Zhongle Xie, Jue Wang, Ke Chen, and Lidan Shou. Continual federated learning based on knowledge distillation. In *Ijcai*, pages 2182–2188, 2022.

[Mallya and Lazebnik, 2017] Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7765–7773, 2017.

[McCloskey and Cohen, 1989] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989.

[McMahan *et al.*, 2016] H. B. McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics*, 2016.

[Orang *et al.*, 2025] Omid Orang, Pedro H Barros, Giulia Zanon de Castro, and Frederico Gadelha Guimarães. An efficient one-shot federated medical imaging via variational inference parametric feature transfer. In *Medical Imaging meets EurIPS: MedEurIPS 2025*, 2025.

[Parisi *et al.*, 2019] German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural networks*, 113:54–71, 2019.

[Rolnick *et al.*, 2019] David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy P. Lillicrap, and Greg Wayne. *Experience replay for continual learning*. Curran Associates Inc., Red Hook, NY, USA, 2019.

[Rusu *et al.*, 2022] Andrei A. Rusu, Neil C. Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks, 2022.

[Shen *et al.*, 2017] Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19(1):221–248, 2017.

[Shoham *et al.*, 2019] Neta Shoham, Tomer Avidor, Aviv Keren, Nadav Israel, Daniel Benditkis, Liron Mor-Yosef, and Itai Zeitak. Overcoming forgetting in federated learning on non-iid data. *arXiv preprint arXiv:1910.07796*, 2019.

[Verma *et al.*, 2023] Tanvi Verma, Liyuan Jin, Jun Zhou, Jia Huang, Mingrui Tan, Benjamin Chen Ming Choong, Ting Fang Tan, Fei Gao, Xinxing Xu, Daniel S Ting, et al. Privacy-preserving continual learning methods for medical image classification: a comparative analysis. *Frontiers in Medicine*, 10:1227515, 2023.

[Wang *et al.*, 2025] Naibo Wang, Yuchen Deng, Shichen Fan, Jianwei Yin, and See-Kiong Ng. Multi-modal one-shot federated ensemble learning for medical data with vision large language model, 2025.

[Yang *et al.*, 2023] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2 - a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1), January 2023.

[Yang *et al.*, 2024] Xin Yang, Hao Yu, Xin Gao, Hao Wang, Junbo Zhang, and Tianrui Li. Federated continual learning via knowledge fusion: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 36(8):3832–3850, 2024.

[Yao and Sun, 2020] Xin Yao and Lifeng Sun. Continual local training for better initialization of federated models. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 1736–1740, 2020.

[Zhang *et al.*, 2021a] Chen Zhang, Yu Xie, Hang Bai, Bin Yu, Weihong Li, and Yuan Gao. A survey on federated learning. *Knowledge-Based Systems*, 2021.

[Zhang *et al.*, 2021b] J Zhang, Chen Chen, Bo Li, Lingjuan Lyu, Shuang Wu, Shouhong Ding, Chunhua Shen, and Chao Wu. Dense: Data-free one-shot federated learning. In *Neural Information Processing Systems*, 2021.

[Zhang *et al.*, 2022] Jie Zhang, Chen Chen, Bo Li, Lingjuan Lyu, Shuang Wu, Shouhong Ding, Chunhua Shen, and Chao Wu. Dense: data-free one-shot federated learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2022.

[Zhou and Wang, 2025] Tianshuo Zhou and Boyuan Wang. Cross paradigm fusion of federated and continual learning on multilayer perceptron mixer architecture for incremental thoracic infection diagnosis. *Scientific Reports*, 15(1):24449, 2025.