Evaluating and Enhancing Data Privacy in Distributed Environments with Non-Parametric Federated Learning using Pre-Trained Foundation Models

Master Thesis

Master of Science in Computing in the Humanities

Hanh Huyen My Nguyen

August 1, 2025

Supervisor:

1st: Prof. Dr. Christian Ledig
2nd: Francesco Di Salvo

Chair of Explainable Machine Learning
Faculty of Information Systems and Applied Computer Sciences
Otto-Friedrich-University Bamberg

Abstract

Deep learning holds great promise for advancing medical image analysis, but access to large and diverse datasets for robust training is often constrained by privacy regulations. Federated Learning (FL) enables collaborative training among distributed institutions without sharing raw data. However, conventional FL approaches, which rely on downstream model sharing, are restricted to specific tasks, incur high communication costs, and remain vulnerable to privacy attacks. We propose a novel FL framework that shifts from model-centric collaboration to privacy-preserving data sharing. By leveraging pre-trained foundation models (FMs), clients extract compact, semantically rich embeddings and share anonymized representations to support local downstream tasks without the need for parametric synchronization.

As anonymization strategies for this framework, we explore (i) a non-parametric approach based on unsupervised clustering and k-anonymity with additive differential privacy (DP) noise (DP-kSame), and (ii) a generative approach using a federated, differentially private Conditional Variational Autoencoder (DP-CVAE) to model a global, privacy-aware data distribution. Both methods enhance client autonomy and support personalized downstream learning with minimal additional training. Validated across multiple medical imaging datasets and feature extractors, our proposed methods outperform traditional FL classifiers while ensuring strong privacy guarantees. Together, they demonstrate the viability of FM-embedding-based data sharing for scalable and secure FL. DP-CVAE achieves the best privacy–utility trade-off, offering superior accuracy, fidelity, adaptability, and robustness against privacy attacks.

The code is available on GitHub[1].

[1] `https://github.com/myng15/privacy-preserving-non-parametric-FL`

Abstract

Deep Learning bietet großes Potenzial für Fortschritte in der medizinischen Bildanalyse, doch der Zugang zu großen und vielfältigen Datensätzen für ein robustes Training wird häufig durch Datenschutzbestimmungen eingeschränkt. Federated Learning (FL) ermöglicht kollaboratives Training zwischen verteilten Institutionen, ohne dass Rohdaten ausgetauscht werden müssen. Herkömmliche FL-Ansätze, die auf dem Austausch von Downstream-Modellen basieren, sind jedoch auf spezifische Aufgaben beschränkt, verursachen hohe Kommunikationskosten und bleiben anfällig für Privacy-Angriffe. Wir schlagen ein neuartiges FL-Framework vor, das von einer modellzentrierten Zusammenarbeit zu einer datenschutzkonformen Datenfreigabe übergeht. Durch die Nutzung vortrainierter Basismodelle (FMs) extrahieren die FL-Clients kompakte, semantisch aussagekräftige Einbettungen und teilen anonymisierte Repräsentationen, um lokale nachgelagerte Aufgaben zu unterstützen, ohne dass eine parametrische Synchronisation erforderlich ist.

Als Anonymisierungsstrategien für dieses Framework untersuchen wir (i) einen nichtparametrischen Ansatz basierend auf unüberwachtem Clustering und k-Anonymität mit additivem Differential Privacy (DP)-Rauschen (DP-kSame) und (ii) einen generativen Ansatz unter Verwendung eines föderierten, differentiell privaten Conditional Variational Autoencoder (DP-CVAE) zur Modellierung einer globalen, datenschutzbewussten Datenverteilung. Beide Methoden stärken die Autonomie der Clients und ermöglichen personalisiertes Downstream-Lernen mit minimalem zusätzlichem Trainingsaufwand. Unsere vorgeschlagenen Methoden wurden anhand mehrerer medizinischer Bilddatensätze und Basismodelle validiert und übertreffen traditionelle FL-Klassifikatoren bei gleichzeitig hohen Datenschutzgarantien. Zusammen zeigen sie die Praktikabilität von FM-Embedding-basiertem Datenaustausch für skalierbares und sicheres FL. DP-CVAE erzielt dabei das beste Privacy-Utility-Verhältnis und bietet höchste Klassifikationsgenauigkeit, Rekonstruktionsgenauigkeit, Anpassungsfähigkeit und Robustheit gegenüber Datenschutzangriffen.

Der Quellcode ist auf GitHub verfügbar[2].

[2] `https://github.com/myng15/privacy-preserving-non-parametric-FL`

Acknowledgements

I would like to express my sincere appreciation to Prof. Dr. Christian Ledig and the entire team at the Chair of Explainable Machine Learning. The courses, projects and further activities I had the opportunity to engage in at the chair have sparked and nurtured my passion for the field of deep learning and Explainable AI and for research in general.

I am deeply indebted to my supervisor, Francesco Di Salvo, whose dedicated guidance, continuous encouragement, and inspiring ideas accompanied me throughout the entire thesis process. His energy and enthusiasm for research have been a constant source of inspiration, helping me develop both the confidence and curiosity to explore far beyond the original scope of this work.

I am especially grateful to Prof. Ledig and my supervisor Francesco for the opportunity to co-author a paper presented at the MICCAI 2025 conference – an achievement that surpassed all my expectations. Their collaboration and mentorship throughout the publication process provided invaluable learning experiences, both academically and personally.

Finally, my heartfelt thanks go to my family and friends for their unconditional support, encouragement, and belief in me. Their presence carried me through the most demanding phases of my studies and thesis writing.

This journey has been a collective effort, and I am forever grateful to everyone who made it possible.

Contents

List of Figures

List of Tables

List of Acronyms

AI	Artificial Intelligence
CVAE	Conditional Variational Autoencoder
DCR	Distance to Closest Record
DINO	Self-DIstillation with NO labels
DP	Differential Privacy
FL	Federated Learning
FM	Foundation Model
GAN	Generative Adversarial Network
GDPR	General Data Protection Regulation
HIPAA	Health Insurance Portability and Accountability Act
IoT	Internet of Things
k-NN	k-Nearest Neighbors
ML	Machine Learning
MLP	Multi-Layer Perceptron
MIA	Membership Inference Attack
NLP	Natural Language Processing
PII	Personally Identifiable Information
PM	Privacy Metric
ROC-AUC	Area Under the Receiver Operating Characteristic Curve
SGD	Stochastic Gradient Descent
VAE	Variational Autoencoder
ViT	Vision Transformer

Notation

Numbers and Arrays

a	A scalar (integer or real)
$\mathbf{a}$	A vector
$\mathbf{A}$	A matrix
$[M]$	A set of indices $\{1, \ldots, M\}$
$\mathbf{I}_n$	Identity matrix with n rows and n columns
$\mathbf{I}$	Identity matrix with dimensionality implied by context
$\mathbf{0}$	The zero vector with dimensionality implied by context

Sets

$\mathbb{A}$	A set
$\mathbb{R}$	The set of real numbers
$\{0, 1\}$	The set containing 0 and 1
$\{0, 1, \ldots, n\}$	The set of all integers between 0 and n
$[a, b]$	The real interval including a and b
$(a, b]$	The real interval excluding a but including b
$\mathbb{A} \backslash \mathbb{B}$	Set subtraction, i.e., the set containing the elements of $\mathbb{A}$ that are not in $\mathbb{B}$

Probability and Information Theory

$P(\mathrm{a})$	A probability distribution over a discrete variable
$P[\cdot]$	A probability of an event
$a \sim P$	Random variable a has distribution P
$\mathbb{E}_{\mathrm{x} \sim P}[f(x)]$	Expectation of $f(x)$ with respect to $P(\mathrm{x})$
$D_{\mathrm{KL}}(P \| Q)$	Kullback-Leibler divergence of P and Q

Functions

$f(\mathbf{x};\theta)$	A function of $\mathbf{x}$ parametrized by θ. (Sometimes we write $f(\mathbf{x})$ and omit the argument θ to lighten notation)
$\ln x$	Natural logarithm of x
$\lfloor a \rfloor$	Floor function (greatest integer $\leq a$)
$\max(a, b)$	Maximum of a and b
$\|\|\mathbf{x}\|\|_2$	ℓ_2-norm of vector $\mathbf{x}$
$\mathbb{1}_{\{\text{condition}\}}$	is 1 if the condition is true, 0 otherwise
Φ	(Feature extraction by) Pre-trained foundation model

Federated Learning

$\mathbb{C}$	A set of unique class labels across clients
M	Number of clients in federation
n_m	Number of samples at client m
$\mathcal{D}_m$	Client m's raw dataset $\{(\mathbf{d}_i^m, y_i^m)\}_{i=1}^{n_m}$
$\mathcal{S}_m$	Client m's embedding dataset $\{(\mathbf{x}_i^m, y_i^m)\}_{i=1}^{n_m}$
$\mathcal{X}_m$	Input space (embedding space) for client m
$\mathcal{Y}_m$	Label space for client m
P_m	Data distribution over $\mathcal{X}_m \times \mathcal{Y}_m$ for client m
$\mathbf{x}'_i$	Anonymized embedding corresponding to the i-th real embedding (DP-kSame)
$\hat{\mathbf{x}}_i$	Synthetic embedding corresponding to the i-th real embedding (DP-CVAE)
$\mathcal{S}'_m$	Anonymized dataset from client m (DP-kSame)
$\hat{\mathcal{S}}_m$	Synthetic dataset generated by client m (DP-CVAE)
$\hat{\mathcal{S}}^{\text{g}}$	Global pooled datastore (DP-kSame)
$\mathcal{D}^{\text{g}}$	Global CVAE decoder (DP-CVAE)

Differential Privacy

(ϵ, δ)	DP privacy parameters
B	Clipping norm bound
σ	Standard deviation of Gaussian noise
Δf	Sensitivity of a function f
$\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$	Gaussian noise distribution

Clustering

$C_m^{(j)}$	Centroid of cluster j at client m
$\tilde{C}_m^{(j)}$	Noised centroid (DP-kSame)
k	Cluster size (for k-anonymity)
J	Number of clusters per client

CVAE

$\mathbf{z}$	Latent variable
$\mathcal{N}(\mathbf{0}, \mathbf{I})$	Standard normal distribution
$q_\phi(\mathbf{z} \mid \mathbf{x}, y)$	Encoder distribution
$p_\theta(\mathbf{x} \mid \mathbf{z}, y)$	Decoder distribution
β	KL divergence weight
$\mathcal{L}_{\text{CVAE}}$	CVAE loss function
$\nabla_{\theta_m, \phi_m} \mathcal{L}_{\text{CVAE}}$	Gradient of CVAE loss w.r.t. client m's decoder (θ_m) and encoder (ϕ_m) parameters

Downstream Models

h_m	Client m's hypothesis/model
θ_m	Parameters of client m's model
$P_{\mathcal{S}_m}(y \mid \mathbf{x})$	Predictive distribution (local data)
$P_{\hat{\mathcal{S}}_m}(y \mid \mathbf{x})$	Predictive distribution (global data)
$\mathbf{x}_{\text{test}}$	Test input embedding (e.g., for evaluation)
$\hat{y}_{\text{test}}$	Predicted label for $\mathbf{x}_{\text{test}}$
λ_m	Interpolation weight for client m

1 Introduction

Data-driven machine learning (ML) methods, particularly deep neural networks, have driven remarkable advancements in various domains, including medical image analysis (Altaf et al., 2019; Litjens et al., 2017). Yet, their clinical adoption remains constrained by data scarcity and stringent privacy regulations (Adnan et al., 2022; Casaletto et al., 2023). Medical datasets are often siloed within institutions, and rare diseases in particular suffer from the lack of diverse, high-quality training datasets (Chen et al., 2024; Rieke et al., 2020). While collaborative data sharing could alleviate these problems, legal frameworks such as HIPAA in the US (Annas, 2003) or GDPR in Europe (Voigt and Bussche, 2017) impose strict limitations on the sharing of personally identifiable health data. This restricts the collective insights necessary for developing robust artificial intelligence (AI) solutions in healthcare.

Federated Learning (FL), first introduced by Google in 2017 (McMahan et al., 2017), offers a promising approach to resolve this fundamental tension between data-driven AI and data privacy protection. It enables collaborative model training without raw data exchange: each participating client (*e.g.,* a hospital) trains a local model on its private data and shares only parameter updates to form a global model (Jere et al., 2021; Khan et al., 2025). However, despite its foundational appeal, FL introduces its own set of challenges. First, (parametric) FL typically requires a shared model architecture, often tied to a single downstream task (*e.g.,* classification or segmentation). This design limits personalization and flexibility when clients have different needs or resources (Marfoq et al., 2022; Tzortzis et al., 2025). Second, privacy remains a concern: even parameter updates can leak sensitive information from the training data through membership inference or model inversion attacks (Casaletto et al., 2023; Jere et al., 2021), and methods that attempt to protect against these attacks often sacrifice model performance (Chen et al., 2022; Rodríguez-Barroso et al., 2023). Third, frequent transmission of large, complex deep learning architectures like Vision Transformers (ViTs) (Dosovitskiy et al., 2021) in FL settings incurs excessively high communication costs.

Our work addresses these challenges by first and foremost pivoting to a latent-space FL approach that facilitates few-shot learning with even basic non-parametric methods like k-means clustering or k-nearest neighbors (k-NN). Instead of training and sharing large parametric models for downstream tasks in the input space, we leverage pre-trained foundation models (FMs) to extract compact, semantically rich feature embeddings (Di Salvo et al., 2024; Doerrich et al., 2024; Li et al., 2024a). Such embeddings are robust to distribution shifts and can be used directly for downstream tasks without fine-tuning, often achieving better performance than representations produced by task-specific models (Caron et al., 2021; Oquab et al., 2024). This makes FM embeddings a strong basis for pursuing various downstream tasks with algorithms as simple as k-NN, motivated by such methods' interpretability and adaptability (Doerrich et al., 2024). Furthermore, we aim at decoupling local training from global aggregation – a direction increasingly advocated in recent FL research as a strategy to improve client autonomy, architecture flexibility, communi-

cation efficiency, and privacy (Chen and Chao, 2022; Gao et al., 2022; Marfoq et al., 2022; Tan et al., 2023; Yurochkin et al., 2019). Non-parametric methods emerge as a natural choice to implement this decoupled update exchange, as they require no shared architectures, avoid gradient leakage, and allow flexible adaptation to different data distributions without the need to retrain upon receiving new data (Imam et al., 2024; Marfoq et al., 2022).

Motivated by these insights, we explore non-parametric alternatives in FL at two levels: first, for global knowledge aggregation and sharing, and second, for the local downstream task (*e.g.,* classification). After extracting robust feature embeddings using a pre-trained FM (*e.g.,* DINOv2), our framework proposes that, instead of training and exchanging local models for global aggregation, clients share anonymized data representations. Specifically, clients can share cluster centroids computed from their embedding dataset via k-means clustering, or apply k-anonymity protection (Sweeney, 2002) to the entire shared embedding dataset, ensuring each shared "representative" is indistinguishable from at least $k-1$ others. Subsequently, for the downstream task, each client uses a non-parametric algorithm such as k-NN on its private embeddings and on the anonymized global embeddings received from the server to predict on test data. The final prediction is then obtained by fusing the local and the global k-NN outputs to balance local specificity and global generalization.

Regarding the global sharing stage, similar prototype- or data-sharing techniques have been explored to improve FL performance, especially in heterogeneous settings (Zhu et al., 2021). Prior works like Tan et al. (2022); Tran et al. (2024); Yoshida et al. (2019); Zhao et al. (2018) have attempted to mitigate the negative effect of data heterogeneity by sharing a limited amount of local data or "knowledge" (*e.g.,* local class prototypes) with the server to regularize the federated training. As Casaletto et al. (2023) suggests, if we can share the capabilities to generate these summary statistics, then we can share knowledge without exposing sensitive patient data. However, these approaches still face the risk of violating FL's privacy-preserving requirements, as clients are still exchanging direct (even if summarized) data representations. To address this, we further apply Differential Privacy (DP) noise (Dwork et al., 2006; Dwork, 2006) to the anonymized embeddings, thereby enforcing a formal privacy guarantee. Figure 1 illustrates our approach to this non-parametric federated data sharing scheme.

While such centroid-based schemes can be straightforward and efficient, they have inherent limitations. They summarize data distributions rather coarsely and cannot generate new, diverse, and task-relevant samples, particularly for underrepresented classes. These shortcomings have prompted us to adopt a different approach in our proposed pipeline: data-sharing via privacy-preserving synthetic data generation. As illustrated in Figure 2, clients collaboratively train a Differentially Private Conditional Variational Autoencoder (DP-CVAE) to model a global, privacy-aware data distribution. Notably, only the decoder weights of the CVAE – which are typically small and contain minimal leakage of raw data – are shared and aggregated. This generative approach therefore substantially improves privacy compliance while

supporting broader downstream tasks (Giuffrè and Shung, 2023; Koetzier et al., 2024; Ktena et al., 2024). Besides, the use of a lightweight generative model like CVAE, operating directly on embeddings instead of raw images, enhances fidelity and robustness of the generated embeddings while reducing communication overhead compared to existing federated generative modeling approaches (see Gargary and De Cristofaro (2024) for a comprehensive review). In particular, training a CVAE on rich FM embeddings further facilitates the accurate representation of feature distributions and the low-cost downstream learning. After generating a globally representative synthetic set *locally* using the shared decoder, a client can train any downstream model (*e.g.,* k-NN or simple linear probes) without further federation. Similar to our exploratory non-parametric pipeline introduced earlier, this synthetic-data-based method nicely decouples data sharing from downstream training.

In summary, our contributions are:

- We shift the FL paradigm from downstream model sharing in the input space to privacy-preserving data sharing in the embedding space, effectively decoupling global representation learning from local downstream tasks for enhanced personalization and generalizability. By leveraging FMs for advanced feature extraction, we enable more efficient and secure data exchanges.
- We introduce two representation sharing approaches: one based on direct, non-parametric k-anonymization mechanisms, and another based on synthetic data generation, leveraging a federated, differentially private generative model (CVAE). By evaluating these methods, we validate the effectiveness of privacy-preserving data sharing via FM embeddings in enhancing the privacy-utility trade-off in FL.
- We empirically demonstrate that our federated generative approach and subsequent (non-parametric) downstream training outperform traditional federated classifiers across multiple medical datasets, achieving superior classification accuracy and privacy-utility trade-offs.
- We show that training a lightweight CVAE on feature embeddings in a federated setting achieves higher-fidelity generation than GAN-based methods while requiring approximately $5\times$ fewer parameters, significantly improving computational efficiency.

The remainder of this work is structured as follows. Section 2 describes background knowledge and reviews relevant prior works on FL and its primary challenges, privacy-preserving techniques, and pre-trained vision FMs. Section 3 elaborates on our proposed methods and supporting theories. Section 4 presents a set of experiments and results on standard medical image datasets (*e.g.,* MedMNIST datasets), including analyses of both theoretical and empirical privacy (versus utility) of our methods. In Section 5, we discuss the key findings and their impact, the limitations of our methodologies, and potential avenues for future research. Finally, we conclude the study in Section 6.

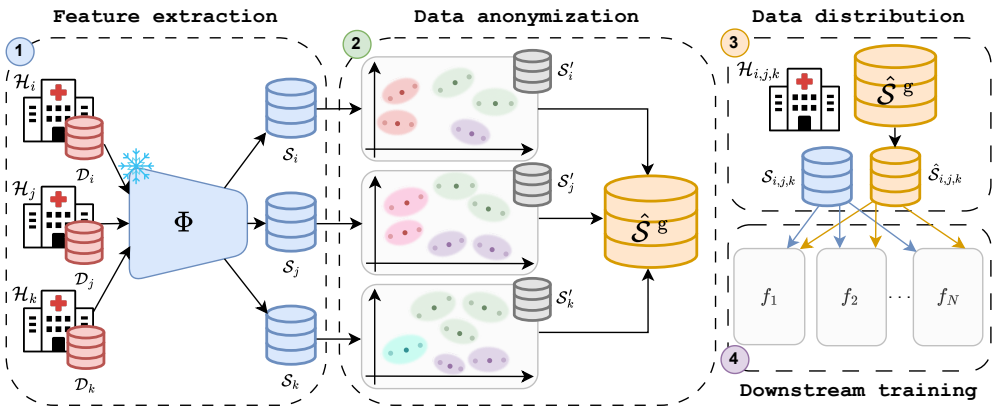

Figure 1: Illustration of our DP-kSame methodology. (1) Each client $\mathcal{H}$ encodes its image-based dataset $\mathcal{D}$ into an embedding-based dataset $\mathcal{S}$ using a large, pre-trained FM Φ. (2) Clients form clusters with k embeddings to ensure k-anonymity ($k = 2$ in this illustration), then replace all embeddings in each cluster with its DP-noised centroid (darker point), and send this k-anonymous dataset $\mathcal{S}'$ into a global datastore $\hat{\mathcal{S}}^{\mathrm{g}}$. (3) A server distributes the global datastore back to each client (excluding the data sent by the respective client). (4) Each client utilizes (real) local and (anonymized) global data for any downstream task f.

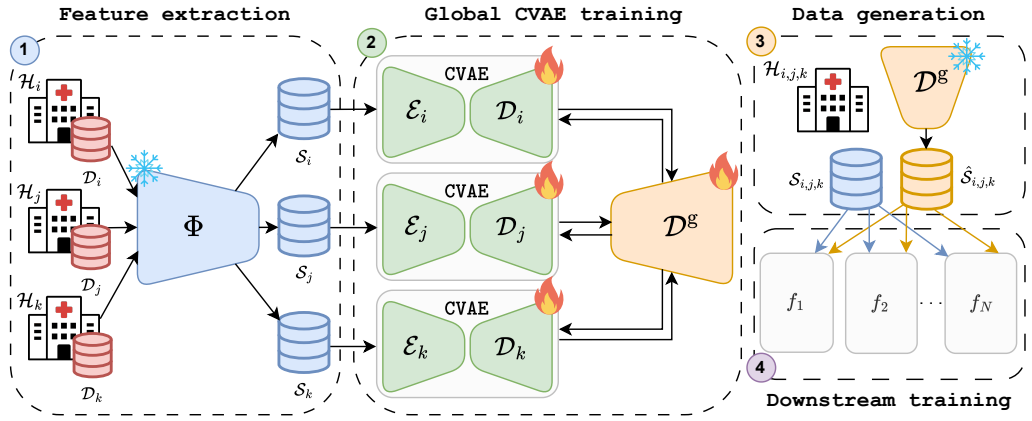

Figure 2: Illustration of our DP-CVAE methodology. (1) Each client $\mathcal{H}$ encodes its image-based dataset $\mathcal{D}$ into an embedding-based dataset $\mathcal{S}$ using a large, pre-trained FM Φ. (2) Clients collaboratively train a DP-CVAE $(\mathcal{E}, \mathcal{D})$ and periodically share decoder weights, which are aggregated into a global decoder $\mathcal{D}^{\mathrm{g}}$. (3) Each client independently generates a synthetic dataset $\hat{\mathcal{S}}$ using the globally trained CVAE decoder $\mathcal{D}^{\mathrm{g}}$, and (4) utilizes (real) local and (synthetic) global data for any downstream task f.

2 Background and Related Work

2.1 Federated Learning: Overview

FL is a collaborative learning paradigm that aims to solve a ML problem using multiple decentralized data sources, referred to as *clients*, while keeping their data local (McMahan et al., 2017; Kairouz et al., 2021; Karimireddy et al., 2022). These clients can be individual mobile or Internet of Things (IoT) devices (*cross-device FL*), or different institutions, organizations or geo-distributed data centers (*cross-silo FL*). Typically, a central server or service orchestrates the training of a shared model, but never sees raw data.

Algorithm 1 `FederatedAveraging` (Adapted from McMahan et al. (2017)).

Input: M participating clients; number of communication rounds T; local mini-batch size n_b; number of local training epochs E; learning rate η
Output: Global model weights $\theta^{(T)}$

Server executes:
Initialize global weights $\theta^{(0)}$
for each round $t = 1, \ldots, T$ **do**
  for each client $m = 1, \ldots, M$ **in parallel do**
    $\theta_m^{(t)} \leftarrow \text{ClientUpdate}(m, \theta^{(t)})$ ▷ start local training with current global weights
  end for
  $s_t \leftarrow \sum_{m=1}^{M} n_m^{\text{train}}$
  $\theta^{(t+1)} \leftarrow \sum_{m=1}^{M} \frac{n_m^{\text{train}}}{s_t} \theta_m^{(t)}$ ▷ update global model with weighted average of client updates
end for

function ClientUpdate(m, θ)
  for each local epoch $e = 1, \ldots, E$ **do**
    for each mini-batch b of size n_b **do**
      $\theta \leftarrow \theta - \eta \nabla f(\theta; b)$ ▷ perform local mini-batch gradient descent
    end for
  end for
  return θ to server
end function

In its standard, parametric form, FL solves a global, distributed optimization problem: it seeks a parameter set $\theta \in \mathbb{R}^d$ that minimizes the total loss across all M participating clients,

$$\min_{\theta} \left\{ f(\theta) := \frac{1}{M} \sum_{m=1}^{M} f_m(\theta) \right\}, \tag{1}$$

where $f_m(\theta) : \mathbb{R}^d \to \mathbb{R}$ denotes the expected loss with respect to client m's local data using the shared parameters θ. A standard algorithm to solve this is FederatedAveraging (or FedAvg), introduced by McMahan et al. (2017). In each communication round t, the server broadcasts the current global model $\theta^{(t)}$ to all M participating clients, who run local steps of stochastic gradient descent (SGD) on their local data to update $\theta^{(t)}$. The server aggregates the resulting weights $\theta_m^{(t)}$ by computing their weighted average as follows to update the global model:

$$\theta^{(t+1)} = \sum_{m=1}^{M} w_m \theta_m^{(t)}, \quad w_m = \frac{n_m^{\text{train}}}{\sum_{m=1}^{M} n_m^{\text{train}}}, \tag{2}$$

where n_m^{train} is the number of training data points at client m. The updated global model is subsequently redistributed to the clients for further iterations. Complete pseudo-code of this process is given in Algorithm 1.

The core principles of FL are data localization and minimization. It is therefore a privacy-by-design framework: local data stays local, and only focused updates (*i.e.,* minimum information necessary for the specific task) are exchanged to achieve the learning objective (Jere et al., 2021; Kairouz et al., 2021). This design not only mitigates many privacy risks faced by traditional, centralized ML, but also reduces network communication overhead compared to raw data transmission (Hallaji et al., 2024; Rauniyar et al., 2024). This efficiency is particularly beneficial for environments with limited bandwidth or high data volumes, such as edge computing or IoT deployments.

In fact, FL was initially introduced with an emphasis on mobile and edge device applications, motivated by the bandwidth and latency limitations inherent in distributed training on such devices (McMahan et al., 2017). Google pioneered cross-device FL with the first production-grade FL system for training language models powering the Gboard mobile keyboard's next-word prediction (McMahan and Ramage, 2017; Jere et al., 2021). Since then, FL has seen widespread adoption across different domains, from edge computing and IoT (Kairouz et al., 2021; Yurdem et al., 2024) to finance (He et al., 2024; Liu et al., 2023; Shi et al., 2023) and natural language processing (NLP) (Du et al., 2023; Lin et al., 2022). Especially, FL has gained a lot of attention in the medical field (Rauniyar et al., 2024; Rieke et al., 2020), since health data is often highly sensitive and its usage is tightly regulated, hindering its centralization for model training (Tan et al., 2023; Rieke et al., 2020). Cross-silo FL is particularly relevant in healthcare, enabling secure collaboration among institutions who are holders of private, siloed datasets. For instance, in joint medical imaging tasks, cross-silo FL can facilitate the training of robust and scalable diagnostic or prognostic models by leveraging medical images from multiple hospitals (Kairouz et al., 2021; Myakala et al., 2024). During the COVID-19 pandemic, researchers have harnessed the potential of FL for a wide range of innovative applications, such as estimating oxygen requirements or detecting lung abnormalities in CT (Liu and Han, 2024; Rauniyar et al., 2024). The ability to leverage decentralized datasets for collective intelligence with minimal data transfer makes the FL

paradigm ideal for scenarios where the need for collaborative insights and timely responses while preserving privacy is paramount.

Despite these benefits, traditional FL (where a parametric model like a neural network is shared for a downstream task) faces significant real-world obstacles, including statistical, computational, communication, and privacy challenges. In the remainder of this subsection, we examine these key challenges of FL and review existing research efforts to address them.

2.2 Heterogeneity Challenges in Federated Learning

In FL, clients often differ both in volume and distribution of their local datasets (*statistical heterogeneity*), and in their storage, computational, and communication resources (*system heterogeneity*) (Marfoq et al., 2022; Saeed et al., 2025; Ye et al., 2023). On the one hand, the decentralized nature of FL allows for scalable and inclusive model training, where devices with varying resources can contribute to and benefit from the collaborative learning. On the other hand, relying on a single global model trained on heterogeneous data across all clients often leads to suboptimal performance and fairness concerns. For example, medical image data from different institutions are typically non-independent and identically distributed (non-IID). Variations in imaging modalities, equipments, and patient populations cause distribution shifts that degrade the performance of the global FL model on individual institutions (Tan et al., 2023; Tzortzis et al., 2025). As summarized by Kairouz et al. (2021), the core challenges of FL arise from unbalanced and non-IID data distributions across numerous unreliable devices with limited bandwidth.

Personalized FL has emerged as a promising solution to the issue of heterogeneity by creating customized models for each client while leveraging the shared knowledge from the others (Chen and Chao, 2022; Morafah et al., 2024). Clustering-based FL methods (*e.g.,* Sattler et al. (2020); Ghosh et al. (2020) partition clients into clusters with similar data distributions, each cluster sharing the same model. Other works (Li and Wang, 2019; Wang et al., 2024; Xu and Fan, 2023) involve distilling the knowledge from a global or historical "teacher" model to clients' "student" models, while allowing federated clients to learn different model architectures suited to their capabilities. Meta-learning-based strategies (Fallah et al., 2020; Liu et al., 2024; Scott et al., 2024; Voleti and Ho, 2024) learn a shared meta-model that enables fast, few-shot local adaptation. Another group of studies (Deng et al., 2020; Mansour et al., 2020; Marfoq et al., 2022) proposes model interpolation techniques, which seek the optimal combination of a global and a local model per client with a controllable mixing weight in order to achieve the best client-specific interpolated model.

While many personalized FL approaches still rely on the joint learning of global and local models, recent works have shown a growing interest in decoupling local model optimization from the global training process. This separation has been demonstrated to resolve more effectively the dilemma in heterogeneous settings about whether to prioritize the learned model's generic performance for future global use

or its personalized performance for individual clients (Chen and Chao, 2022). Additionally, decoupled FL methods promise enhanced privacy when personalization occurs separately from federated communication and clients do not need to share information about their local models with the server (Marfoq et al., 2022). Methods like FedPer (Arivazhagan et al., 2019), FedRep (Collins et al., 2021), and pFedGP (Achituve et al., 2021) approach a global-local decoupling by separating model layers into a globally trained base (for representation learning) and a locally trained head (for personalized learning). However, the global representation learned by the base layers is still affected by local refinement (*i.e.,* by the gradient updates from the local heads). In contrast, kNN-Per (Marfoq et al., 2022) offers a more pronounced decoupling between local model training and the federated learning of a shared representation. It interpolates predictions from a global parametric model with local k-NN outputs computed from client-specific datastores. Similarly, FedMeS (Xie et al., 2024) leverages local memory and kNN-based inference for personalized federated *continual* learning. In both methods, however, the globally learned model remains tied to a single, specific downstream task (*e.g.,* classification), which can limit their generalizability across different applications.

Memorization-based, decoupled FL techniques like kNN-Per and FedMeS motivate us to explore k-NN for producing client-specific predictions in our interpolation-based FL pipeline as we strive towards a clearer separation between global and local learning to improve personalization and privacy. Unlike parametric models, non-parametric methods like k-NN do not assume a fixed functional form for the model, as illustrated in Figure 3. They discover patterns directly from the data and enable a potentially better representation of complex data distributions, thus increasing prediction accuracy (Imam et al., 2024). Importantly, non-parametric algorithms offer increased flexibility and adaptability to heterogeneous data while not requiring continuous joint optimization or parameter synchronization between global and local models. kNN-Per, for instance, allows quick adaptation to changes in a client's data distribution simply by updating the local datastore without having to retrain the global model (Marfoq et al., 2022).

For producing global-side predictions (based on globally aggregated knowledge), our framework involves clients sharing data-based representatives instead of model updates for global aggregation. One method we explore is to share data based on centroids of clusters formed in the local data via k-means clustering. In general, the idea of sharing a small, globally balanced dataset and using it in conjunction with private data to improve FL performance in non-IID settings has appeared in multiple studies (Zhao et al., 2018; Zhu et al., 2021; Yoshida et al., 2019). Closer to our method are prototype-based FL techniques (Tan et al., 2022; Tran et al., 2024; Voleti and Ho, 2024) - another group of personalized FL strategies. They address heterogeneity by letting clients share abstract data representations (*e.g.,* class means or cluster centroids) instead of gradients, and use the globally aggregated prototypes to refine their own models. These prototypes provide information about local data distributions while minimizing direct data exposure, reducing communication costs and mitigating privacy risks. Like meta-learning-based techniques, prototype-

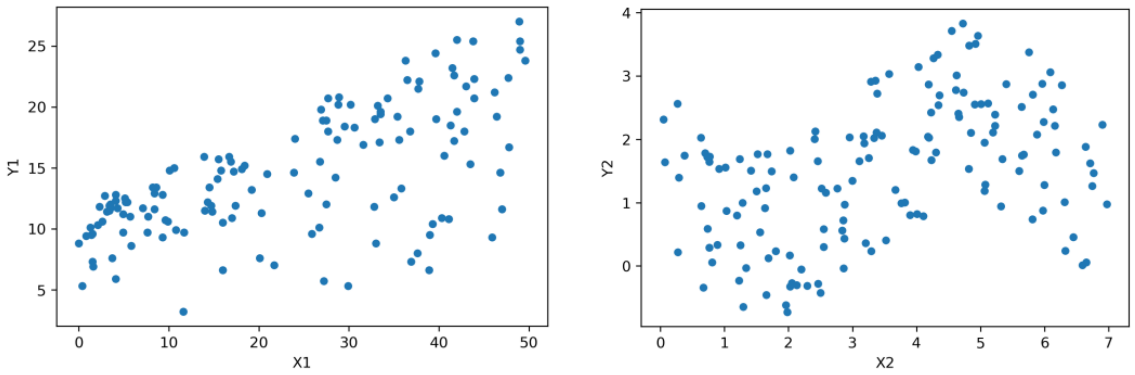

Figure 3: The attributes X1 and Y1 tend to follow a linear relationship and can be modeled using a linear parametric technique, while the attributes X2 and Y2 does not follow any known or linear distribution; therefore, a nonparametric technique would be a better choice. This figure is reproduced from Imam et al. (2024).

based approaches allow clients to use their own model architectures with different input/output spaces, breaking the assumption of federation-wide identical architectures made by most parametric FL methods. Our work refines this prototype-sharing strategy by sharing multiple centroids of arbitrary clustered embeddings instead of a single centroid per class, preserving richer diversity and client-specific knowledge while preventing class-related information leakage. Another difference is that we collect these local prototypes and distribute them back to clients "as they are" without aggregating them into global prototypes to maintain each client's unique contribution. To enhance the privacy of these shared data representatives, different privacy mechanisms are considered, which we will discuss in the following subsection.

2.3 Privacy Challenges in Federated Learning

While FL aims to preserve privacy by keeping raw data local, it does not by itself provide a strong privacy guarantee (Adnan et al., 2022; Ziller et al., 2021). An adversary observing model updates (gradients or weights) can still infer sensitive information about the clients' data. For instance, studies in medical imaging have demonstrated that shared gradients from models trained without additional privacy-preserving techniques can be reverse-engineered to reconstruct images that reveal a patient's identity or medical condition (Kaissis et al., 2021; Ziller et al., 2021, 2024). These *model inversion attacks* are a type of *reconstruction attack*, which is just one of many adversarial threats against FL.

2.3.1 Privacy Threats

Rodríguez-Barroso et al. (2023) and Jere et al. (2021) describe similar taxonomies of adversarial attacks in FL, differentiating between *data privacy attacks* (inferring

sensitive information from the learning process), and *model performance attacks* (modifying the behavior of the federated model, thereby degrading its performance).

Data Privacy Attacks As mentioned above, one type of data privacy attacks is *reconstruction attacks*, which aim to recover the training data of a client participating in a FL task from the exchanged model updates.

Another sub-category is *attribute inference attacks*, which aim to extract whether a property of a client or the entire FL participant population, potentially uncorrelated with the main task, is present in the FL model. For instance, if a ML model's objective is face detection, an attacker might infer if there are training images featuring blue-eyed faces, an attribute not expected to be shared. Melis et al. (2019) observed that modern deep learning models develop separate internal representations of various features, some of which are independent of the main task. These "unintended" features can leak information about participants' training data. An active adversary can manipulate the joint model into learning to better separate the features of interest, thereby extracting even more information.

Membership inference attacks (MIAs) form another important group of data privacy attacks. Their main objective in FL is to determine whether a specific data record (available to the adversary) was part of a client's training dataset. For example, Melis et al. (2019) achieved high accuracy in inferring when a certain person first appeared in photos used to train a binary gender classifier. In a synthetic data context, an MIA is when an attacker tries to identify if certain real records have been used to train the synthetic data generation algorithm. This is a privacy risk since, for example, if the synthetic dataset is about breast cancer, then the attacker can deduce if the person they found has breast cancer (Steier et al., 2025).

Model Performance Attacks Adversarial attacks on the federated model aim to compromise the joint model performance rather than revealing data. Frequently, these attacks are carried out from the client side, leveraging the fact that FL clients can send poisoned updates, while the server cannot inspect the training data stored on the clients.

This group can be categorized into *data poisoning attacks* and *model poisoning attacks*. In data poisoning, the attacker has access to the training data of one or more clients and is able to modify it, *e.g.,* by shuffling labels at random. Model poisoning covers a broad range of methods to manipulate the FL training process, such as poisoning local model updates or altering training rules.

2.3.2 Privacy-Enhancing Techniques

In response to these threats, we propose first and foremost a shift from sharing parameter- or gradient-based updates - typically used for direct training of a shared downstream model - to exchanging anonymized, task-agnostic data representations in the embedding space. A key enabler of this approach is the use of *pre-trained FMs*, which we will review in Section 2.5. By utilizing these powerful feature extractors,

our FL pipeline can achieve strong downstream performance without requiring access to local downstream models. This strengthens defenses against gradient-based privacy attacks that exploit the implicit leakage of sensitive information from raw local data through shared model updates. However, FMs alone do not provide a formal privacy guarantee, and additional privacy-enhancing procedures for the shared feature embeddings are needed.

Simple perturbation methods, such as adding noise to local data or shared model parameters (Chen et al., 2022; Rodríguez-Barroso et al., 2023), can reduce information leakage, but often diminish data utility. Cryptographic methods like *secure multiparty computation* or *homomorphic encryption* allow clients to encrypt model updates before global aggregation, ensuring no party learns anything about the other parties' contributions. Despite offering strong privacy, these approaches typically impose high computational and communication overhead, limiting their practicality for large-scale FL with resource-constrained participants (Chen et al., 2022; Krishnamoorthy, 2025).

Data anonymization techniques, such as k-anonymity (Sweeney, 2002), ℓ-diversity (Machanavajjhala et al., 2007), and t-closeness (Li et al., 2007), aim to obscure or remove sensitive personal attributes to prevent individual identification. k-anonymity ensures each record in a dataset is indistinguishable from at least $k-1$ others. ℓ-diversity further requires diversity in sensitive attributes within these indistinguishable groups, while t-closeness mandates that the distribution of sensitive attributes within a group closely matches the overall dataset distribution. Building on k-anonymity, the k-Same algorithm (Newton et al., 2005) achieves anonymity while minimizing information loss by replacing each data point within disjoint clusters of k similar data points with a single surrogate, typically the cluster centroid.

Similarly, prototype-based FL approaches, such as FedProto (Tan et al., 2022) and FedNTProto (Tran et al., 2024), exchange "data representatives" instead of model parameters. These methods inherently provide a degree of privacy by hiding identifiable individuals behind averaged representations.

In practice, however, these anonymization methods come with critical limitations. Besides potential losses in data quality, the anonymized data remain vulnerable to de-anonymization, for instance, by cross-referencing them with some public dataset (Chen et al., 2022). A classic example is the Netflix case, where Narayanan and Shmatikov (2008) demonstrated that anonymized movie ratings could be accurately linked to users' public IMDb profiles for re-identification. The GDPR specifies that anonymized datasets that can be re-identified with "reasonable effort" are still considered personal data (Curelariu and Lodie, 2024). Additionally, although techniques like k-anonymity are effective for simple datasets, their practicability for anonymizing high-dimensional, diverse data is questionable (Abadi et al., 2016; Aggarwal, 2005; Brickell and Shmatikov, 2008; Narayanan and Shmatikov, 2008). They may also not adequately protect against "singling out" attacks, which the GDPR defines as an adversary's ability to find a predicate that matches exactly one individual in a dataset without knowing their identity (Cohen and Nissim, 2020; Nissim,

2021). Consequently, data anonymization often requires combination with other privacy-enhancing techniques (Chen et al., 2022).

2.3.3 Differential Privacy

A standard and more rigorous approach to enforce formal privacy guarantees is Differential Privacy (DP), as recommended by numerous recent works (Abadi et al., 2016; Kairouz et al., 2021; Palle, 2018; Ziller et al., 2021, 2024). Introduced by Dwork et al. (2006), DP is a mathematical framework for quantifying the privacy provided by a protocol. It ensures that the contribution of any single individual's data point is statistically indistinguishable, meaning the privacy risk to an individual does not significantly change whether their data is included or not in an analysis. Formally, a mechanism $\mathcal{M}$ applied to the private dataset $\mathcal{D}$ is (ϵ, δ)-differentially private if, for any two neighboring datasets $\mathcal{D}$ and $\mathcal{D}'$ differing by at most one sample, and for any possible output $\mathcal{S} \subseteq \text{Range}(\mathcal{M})$:

$$P[\mathcal{M}(\mathcal{D}) \in \mathcal{S}] \leq e^{\epsilon} P[\mathcal{M}(\mathcal{D}') \in \mathcal{S}] + \delta, \quad \epsilon > 0, \delta \in [0, 1) \tag{3}$$

Here, ϵ measures the maximum privacy loss permitted, *i.e.,* our privacy budget. δ bounds the probability of exceeding the privacy budget given by ϵ, so that we can ensure that with probability $1 - \delta$, the privacy loss will not be greater than ϵ. Lower values of ϵ and δ imply stronger privacy. This provides a formal upper bound on an attacker's ability to reconstruct or re-identify specific data points. A standard way to achieve (ϵ, δ)-DP for a function is to add an amount of noise to the function's output. The output of a function f applied to the dataset $\mathcal{D}$ perturbed by noise η is therefore:

$$\mathcal{M}(\mathcal{D}) := f(\mathcal{D}) + \eta \tag{4}$$

This additive noise is determined by (ϵ, δ) as well as the function's sensitivity, defined as

$$\Delta f = \max_{\mathcal{D}, \mathcal{D}'} \|f(\mathcal{D}) - f(\mathcal{D}')\| . \tag{5}$$

DP views privacy as a quantifiable resource, which is used up as information is extracted from a dataset. The goal of private data analysis is, therefore, to extract as much useful information as possible while consuming the least privacy (Adnan et al., 2022). This privacy budget can be calibrated according to policy requirements (Ziller et al., 2024). DP has also been shown to better prevent the aforementioned "singling out" attacks compared to k-anonymity (Cohen and Nissim, 2020; Nissim, 2021).

The use of DP in (parametric) FL was introduced at Google (Abadi et al., 2016) as an extension to the FL paradigm proposed by McMahan et al. (2017). This process is termed *Differentially Private Stochastic Gradient Descent* (DP-SGD). Each client clips the norm of *per-sample* gradients to limit the influence of a single data point, and adds noise to those gradients before sending updates to the server. This ensures

that even with knowledge of the learning algorithm, an adversary cannot confidently infer any single data point. DP defends against various common attacks in FL, such as gradient-based reconstruction attacks (see Figure 4). In our non-parametric approach using k-Same for anonymizing shared data, we apply DP noise to the data instead of gradients, establishing a method that will be detailed in Section 3.2.1.

2.3.4 Privacy-Preserving Synthetic Data Sharing

A promising alternative privacy strategy involves generating *synthetic data* that mimic the real data distribution. Downstream tasks are then trained on synthetic samples, which ideally preserve the utility of real data while containing no exact identifiers. This approach can reduce privacy risks compared to direct data sharing and increase utility compared to traditional anonymization methods (Kaabachi et al., 2025). In the medical domain, several studies have demonstrated the usefulness of synthetic data as proxy for real data (Wang et al., 2019; Azizi et al., 2021; Beaulieu-Jones et al., 2019; Choi et al., 2017), or in augmenting the volume and variability of available data (Jiang et al., 2021; Sufi, 2024).

Data synthesis can be achieved using a wide range of generative models. Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) train two competing neural networks to produce realistic data: a discriminator, which learns to distinguish between real and synthetic data, and a generator, which tries to create synthetic data that is indistinguishable from real data so that it can "fool" the discriminator. Variational Autoencoders (VAEs) (Kingma and Welling, 2014) are another class of neural network models used to generate synthetic data. Like all standard autoencoders, VAEs consists of an encoder that transforms input data into a compact latent representation and a decoder that reconstructs the input from this representation. However, instead of encoding inputs as fixed points, the VAE encoder maps each input into a probability distribution in the latent space – parameterized by a mean and a variance vector. The decoder learns to sample from these distributions to generate new data that reflect the statistical properties of the original dataset.

Of particular interest are the conditional variants of these generative models. Conditional GANs (CGANs) (Mirza and Osindero, 2014) and Conditional VAEs (CVAEs) (Sohn et al., 2015) allow for generating data conditioned on specific information, such as class labels, which is useful for mimicking original class distributions and even augmenting underrepresented classes. For instance, conditional synthetic data generation has been instrumental in augmenting data volume for imaging studies during the COVID-19 pandemic (Jiang et al., 2021) and improving the accuracy of COVID-19 detection by classifying patients based on chest CT scans (Das et al., 2022). Especically, a recent work has demonstrated that generating synthetic feature embeddings using a CVAE preserved classification performance comparable to real embeddings, while enhancing data privacy (Di Salvo et al., 2024). Training a CVAE on feature embeddings rather than raw images further allows to better capture feature distributions (*e.g.,* than k-Same), making it less susceptible to fidelity degradation.

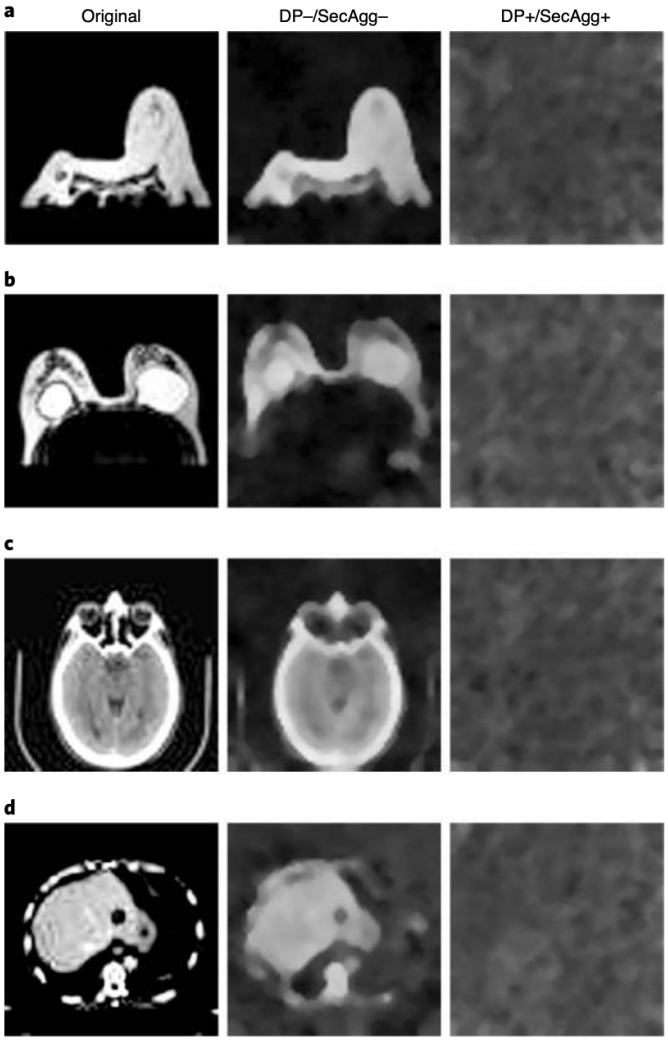

Figure 4: Overview of the gradient-based privacy attacks against medical image analysis using the MedNIST dataset in a variety of scenarios. The original image is shown ("Original") alongside the reconstruction results from a model trained without secure aggregation or DP (DP–/SecAgg–) as well as a model trained with DP and SecAgg (DP+/SecAgg+). In every case, the attack reveals confidential information about the patient when the model is trained without privacy-enhancing techniques. In (a), Breast MRI reveals absence of the right breast, likely due to operative removal due to breast cancer. In (b), Breast MRI reveals breast implants. Both (a) and (b) also allow assumptions about the patient's sex. (c) shows cranial computed tomography image at the level of the nose. Facial contours reconstructed from such images can lead to personal identification (Parks and Monson, 2017; Schwarz et al., 2019). (d) shows Abdominal CT at the level of the liver, which allows visualization of a hypodense lesion in the left liver lobe in the reconstructed image. In every case, using DP thwarts the attack, disallowing any usable image features from being visualized. This figure is reproduced from Kaissis et al. (2021).

In a FL setting, one can imagine each client training a local generator on its data or using a shared global generative model. Alternatively, they can train together a *federated generative model* to capture cross-client data variations while preserving local data privacy. This particular setting has been explored in many recent works (Gargary and De Cristofaro, 2024). A wide range of generative models from GANs to diffusion models (Ho et al., 2020) and Large Language Models (LLMs) (Brown et al., 2020) have also been deployed in federated settings (Zhang et al., 2024; Sufi, 2024). There is also medGAN (Choi et al., 2017), which combines an autoencoder and generative adversarial networks to generate realistic synthetic patient records. These architectures may, however, be unnecessarily heavy for synthesizing embeddings in the latent space of powerful pre-trained FMs. VAEs and CVAEs are more lightweight generative models that have also been explored in FL. However, prior work has primarily applied them to simpler generative tasks, such as MNIST-like datasets (Pfitzner and Arnrich, 2022) and sensor data (Kaspour and Yassine, 2023), or jointly trained the generative model with a downstream classifier (Chen et al., 2023a), limiting their adaptability. This motivates our extension of CVAE to a federated setting, where, as illustrated in Figure 2, clients collaboratively train a global CVAE decoder and then use it locally to generate synthetic embeddings from medical datasets for downstream tasks. Unlike prior FL settings, our approach decouples generative modeling from task-specific constraints. It offers similar privacy and task flexibility as the decoupling of global and local learning discussed in Section 2.2.

Since the use of synthetic data alone is insufficient to guarantee a formal level of privacy, it needs to be used in conjunction with a rigorous privacy mechanism such as DP, carefully designed to keep data unidentifiable without compromising the realism and diversity of the synthesized samples. Differentially private learning of generative models has been studied mostly under the GAN framework, including techniques like DPGAN (Beaulieu-Jones et al., 2019; Cao et al., 2021; Frigerio et al., 2019; Xie et al., 2018), DP-CGAN (Sun et al., 2023; Torkzadehmahani et al., 2019), and PATE-GAN (Jordon et al., 2018). In contrast, recent works on generating data with a CVAE for privacy-preserving data sharing such as Di Salvo et al. (2024) did not include a formal privacy guarantee analysis. Therefore, in this work, we integrate DP into the FL training of both CGAN and CVAE. In this way, our DP-CVAE approach directly addresses the limitations of previous works, ensuring that the synthetic features have provable (ϵ, δ)-differential privacy.

2.3.5 Trade-off between Privacy and Utility

A common theme across all privacy strategies in FL is the inherent trade-off between the strength of privacy guarantees provided by a mechanism and the utility of its output, whether it be model accuracy or data fidelity (Chen et al., 2022; Kairouz et al., 2021; Rodríguez-Barroso et al., 2023).

For instance, adding more DP noise for stronger privacy (smaller ϵ and δ) can degrade model performance. Similarly, synthetic data generation must balance realism with the risk of memorization or distortion of the original data distribution. This

often leads to difficult choices for an appropriate privacy budget. Higher budgets correspond to less privacy protection and thus an increased risk of successful attacks, while lower budgets limit the information available for training. Therefore, striking a balance between privacy and utility is often a fundamental issue when applying privacy-preserving techniques in the FL framework.

2.4 Communication and Computational Challenges in Federated Learning

FL's decentralized training paradigm introduces critical communication and computational challenges. Large-scale FL systems with complex model architectures typically require frequent synchronization of millions of parameters. This places considerable strain on network bandwidth and results in high latency and energy consumption, especially on resource-limited clients (Myakala et al., 2024; Rauniyar et al., 2024; Wu et al., 2022). In addition, unreliable or inconsistent connectivity between participating institutions can also hinder the FL process. Heterogeneous clients with vastly different compute capabilities, ranging from high-performance servers to low-power edge devices, introduce further complexity, as uniform training of large models across all participants becomes impractical (Marfoq et al., 2022; Mao et al., 2022). To enable scalable and inclusive FL, communication- and computation-efficient strategies must be developed without sacrificing downstream performance (Myakala et al., 2024; Tran et al., 2024).

One promising direction is to move away from full model synchronization and instead adopt non-parametric data sharing approaches. Prior work (Tan et al., 2022; Tran et al., 2024) has shown that communicating compact embedding-level class prototypes, rather than entire model weights, can reduce communication costs by orders of magnitude while preserving classification accuracy. These prototype-based approaches also offer inherent privacy advantages, as sharing abstract data representations poses lower privacy risks than transmitting raw data or full gradient updates.

Non-parametric models such as k-NN, used in methods like kNN-Per (Marfoq et al., 2022), offer further benefits in terms of computational and communication efficiency. These models allow for fast, retraining-free adaptation to distribution shifts, and naturally support personalized decision-making. Importantly, they decouple global and local modeling, thereby relieving stronger clients from the need to align their model updates to weaker peers. This allows each participant to independently adapt based on its specific compute and storage capacity.

While kNN-Per still depends on a globally trained model to provide local data embeddings used by local k-NN classifiers, we adopt a fixed global backbone for local feature extraction, leaving clients to perform lightweight downstream adaptation. This decoupling of representation learning from downstream tasks not only reduces communication and computation demands, but also supports better privacy preservation by minimizing the private information exchanged across clients. Indeed,

strategies addressing communication and computation bottlenecks in FL must be designed to not compromise privacy, and, ideally, to enhance the protection and reduce attack surfaces (Jere et al., 2021).

To mitigate computational and communication costs, FL research also explores the use of lightweight architectures. For example, we can employ shallow linear models instead of deep neural networks for downstream classification, and opt for computationally efficient VAEs rather than heavy GAN- or diffusion-based models for federated generative modeling. However, such simplifications often come at the cost of performance and robustness when used in isolation. To reconcile efficiency with strong performance, we leverage high-quality feature representations from powerful pre-trained FMs. This strategy underpins our approach throughout this work and is introduced in the following section.

2.5 Pre-trained Foundation Models

FMs have emerged as a transformative development in modern ML, especially in the field of NLP and, more recently, computer vision (Babé et al., 2025; Zhang et al., 2025). These models are typically large-scale neural networks, pre-trained on vast, diverse datasets using self-supervised or weakly supervised objectives. Notable vision FMs include CLIP (Radford et al., 2021), DINO (Caron et al., 2021), and DINOv2 (Oquab et al., 2024), which leverage architectures like ViTs to learn semantically rich, general-purpose feature representations. Unlike traditional models trained for specific tasks, pre-trained FMs are designed to be task-agnostic and adaptable. These properties make FMs particularly attractive for FL, where heterogeneity, privacy, and communication efficiency are major challenges.

FMs such as DINOv2 are trained on hundreds of millions of images from diverse domains, producing embeddings that generalize well across input corruptions, distribution shifts, and downstream tasks (Kazmierczak et al., 2025; Paul and Chen, 2022). These embeddings capture both fine-grained textures and high-level semantics, enabling even simple classifiers such as k-NN to perform competitively without task-specific fine-tuning. For instance, Caron et al. (2021) show that a 20-NN classifier operating on DINO features performs on par with trained linear models. This robustness and generalization in representation learning are valuable in FL applications, especially in healthcare, where statistical and system heterogeneity often render end-to-end federated training of deep architectures impractical.

This synergy with non-parametric methods like k-NN is a particularly promising aspect of FM embeddings. Non-parametric classifiers are often fast and light to deploy, and require no domain adaptation, making them practical in resource-constrained environments. They are also easy to update and maintain, allowing data points to be added or removed from a datastore without retraining the model. Doerrich et al. (2024) emphasize this flexibility as essential for supporting privacy-preserving ML, including compliance with regulations such as the *right to erasure* (Article 17 of the GDPR). They further demonstrate that combining k-NN with the robust and discriminative feature spaces of vision FMs improves interpretability and adaptability,

particularly in medical image analysis. The strong performance of k-NN classifiers on DINO embeddings without fine-tuning, as shown in Caron et al. (2021), indicates that representations provided by such FM architectures are well-suited for non-parametric evaluation. Li et al. (2024a) also note that the pre-learned, transferable knowledge in FMs and its few-shot learning capabilities can accelerate the FL process and reduce the need for extensive retraining. In this work, we evaluate k-NN classification alongside linear models to study the downstream adaptability of FM embeddings in our FL settings, especially under privacy-aware constraints.

Prior work has also highlighted the inherent privacy and communication benefits of operating in the latent space of FMs. For example, Di Salvo et al. (2024) emphasize the compactness (low dimensionality) and reduced redundancy of these representations, making them ideal for data exchange in distributed learning systems. Importantly, these embeddings are also more privacy-preserving than raw data. Krishnamoorthy (2025), for instance, argue that it is difficult to reconstruct original inputs from latent representations resulting from complex, nonlinear transformations without access to the encoding process.

In summary, the past works outlined above motivated our use of pre-trained vision FMs (particularly the DINOv2 backbone) in this work. They play a central role in our proposed methods, enabling a shift from model-centric FL to lightweight, embedding-based collaboration. By providing compact, robust, and task-agnostic features, FMs facilitate efficient data sharing and adaptable, few-shot downstream learning. At the same time, they reduce communication and computational overhead while enhancing privacy. These advantages are fundamental to our pipelines and establish a new class of collaborative ML systems that rely on efficient, adaptable, and privacy-aware data sharing in the latent space.

3 Methods

3.1 Overview and Notations

In this work, we explore two novel FL approaches that decouple local downstream tasks from (parametric) global learning, aiming to address key FL challenges such as data privacy and communication overhead. The first approach employs non-parametric clustering (via k-means and k-Same algorithms) combined with DP to anonymize client data prior to sharing. This method is referred to as *DP-kSame*. The second approach instead leverages federated, differentially private generative modeling using a CVAE to facilitate knowledge transfer through synthetic data. We refer to this method as *DP-CVAE*. In both methods, clients utilize both private local data and globally shared knowledge to perform personalized downstream tasks. In particular, for image classification, we evaluate both non-parametric models (*e.g.,* k-NN) and parametric ones (*e.g.,* linear classifiers).

Consider a federation of M clients with a total of $n = \sum_{m=1}^{M} n_m$ samples belonging to a set $\mathbb{C} \subset \mathbb{N}$ of unique classes. Each client $m \in [M] := \{1, \ldots, M\}$ has a private image-based dataset $\mathcal{D}_m := \{(\mathbf{d}_i^m, y_i^m)\}_{i=1}^{n_m}$, where $\mathbf{d}_i^m$ denotes the i-th raw image at client m and $y_i^m \in \mathbb{C}$ is its label. For simplicity, we assume all M clients participate in every round of the FL process. Both DP-kSame and DP-CVAE follow a shared pipeline consisting of four main stages:

1. **Feature Extraction:** Each client encodes their image data using a shared, frozen FM Φ, yielding an embedding-based dataset $\mathcal{S}_m := \{(\mathbf{x}_i^m, y_i^m)\}_{i=1}^{n_m}$ where $\mathbf{x}_i^m \in \mathcal{X}_m \subseteq \mathbb{R}^d$ is the feature embedding of image $\mathbf{d}_i^m$ in the representation space of Φ. For brevity, when referring to a single client, we drop the superscript and write $\mathcal{S}_m := \{(\mathbf{x}_i, y_i)\}_{i=1}^{n_m}$. Under data heterogeneity, each $\mathcal{S}_m$ is drawn i.i.d. from a client-specific distribution P_m; for instance, clients may hold disjoint or imbalanced subsets of classes in $\mathbb{C}$.

2. **Global Knowledge Aggregation:**

   (a) **DP-kSame:** Each client m clusters its embedding dataset $\mathcal{S}_m$ using the k-means algorithm, forming groups of k samples to satisfy the requirements of k-anonymity. Note that the k in k-means denotes the number of clusters and is distinct from the k used in this work to indicate the number of samples per cluster. Gaussian DP noise is added to each cluster centroid, which is then used to replace all embeddings within the corresponding cluster (as in the k-Same algorithm). This yields an anonymized dataset $\mathcal{S}'_m$ that satisfies both k-anonymity and (ϵ, δ)-DP. Each client transmits its anonymized embeddings $\mathcal{S}'_m$ to the server.

   (b) **DP-CVAE:** Each client trains a lightweight CVAE on $\mathcal{S}_m$ and periodically shares their decoder weights. A DP mechanism is applied during

local training to ensure each shared decoder update is differentially private. The server aggregates these updates (via FedAvg) into a global decoder $\mathcal{D}^{\mathrm{g}}$.

3. **Global Knowledge Distribution:**

   (a) **DP-kSame:** The server merges $\mathcal{S}'_m$ from all clients into a global datastore $\hat{\mathcal{S}}^{\mathrm{g}}$ and distributes it so that each client m receives $\hat{\mathcal{S}}_m = \hat{\mathcal{S}}^{\mathrm{g}} \setminus \mathcal{S}'_m$ (*i.e.,* all anonymized data from other clients).

   (b) **DP-CVAE:** Using the global CVAE decoder $\mathcal{D}^{\mathrm{g}}$, each client m independently generates a synthetic dataset $\hat{\mathcal{S}}_m$ by sampling embeddings conditioned on a chosen class distribution tailored to its downstream needs. The generated set approximates the global data distribution while remaining privacy-preserving.

4. **Downstream Training:** Each client performs downstream training on its real local dataset $\mathcal{S}_m$ and the global dataset $\hat{\mathcal{S}}_m$ it receives (either anonymized or synthetic). To preserve personalization, models are trained separately on local and global data, and their predictions are combined via a weighted interpolation. A weighting parameter λ_m, tunable via local validation, controls the contribution of each model to the final prediction, enabling a balance between local specificity and global generalization.

The complete workflows of DP-kSame and DP-CVAE are illustrated in Figure 1 and Figure 2 respectively. Both methods leverage the semantic richness of FM embeddings and DP mechanisms to support secure, flexible, and data-efficient knowledge sharing in FL. Notably, the generative approach in DP-CVAE can help mitigate long-tailed distribution issues by augmenting underrepresented classes. The subsequent subsections describe each method in further detail.

3.2 Privacy-Preserving Federated Data Sharing

3.2.1 Non-parametric Data Anonymization via DP-kSame

To avoid sharing model parameters, we adopt a non-parametric prototype-based sharing technique. In existing prototype-based FL methods such as FedProto (Tan et al., 2022) and FedNTProto (Tran et al., 2024), each client m computes a class prototype $C_m^{(c)} = \frac{1}{n_m^{(c)}} \sum_{(\mathbf{x},y)\in\mathcal{S}_m^{(c)}} \mathbf{x}$ to represent class $c \in \mathbb{C}$. These local prototypes are aggregated into global prototypes $\bar{C}^{(c)} = \frac{1}{M_c} \sum_{m=1}^{M_c} C_m^{(c)}$, where M_c is the number of clients with data from class c. The global prototypes are shared back with clients and used as alignment targets: for example, a regularization term penalizes the distance between $C_m^{(c)}$ and $\bar{C}^{(c)}$ in each client m's loss function. Similar to our pipeline, these methods compute prototypes from the set of feature vectors $\mathcal{S}_m$ extracted from raw inputs using a feature extractor f_m . Sharing only one aggregated vector per class

reduces both privacy leakage and communication overhead compared to transmitting a full set of raw data.

We extend these prototype-based methods by: (i) using a powerful FM Φ as a shared feature extractor across clients instead of local embedding functions (*e.g.,* representation layers) f_m, and (ii) sharing a k-anonymous dataset $\mathcal{S}'_m$ of size n_m, instead of a single prototype per class. Following the k-Same algorithm (Newton et al., 2005), each client m clusters its dataset $\mathcal{S}_m$ into groups of size k (ignoring original labels), computes the centroid of each cluster, and assigns a label to the centroid via majority voting. This unsupervised clustering offers a privacy benefit: instead of using true class labels, which might leak sensitive information – especially in the case of outliers, the method only shares the derived labels. Each cluster centroid then replaces all its member embeddings, producing an anonymized dataset $\mathcal{S}'_m$ that contains k identical copies of each centroid. This ensures k-anonymity: each data point in $\mathcal{S}'_m$ could be any of the k originals in the cluster, making individual re-identification unlikely beyond a probability of $1/k$ (Sweeney, 2002). Furthermore, unlike methods that only share class prototypes or cluster centroids, we share the full anonymized dataset as a surrogate for real data, mitigating information loss and enabling richer downstream use (Meden et al., 2018).

Enhancing Privacy with DP To strengthen privacy, we add DP noise to the cluster centroids before applying the k-Same replacement. This addition is motivated by Garg and Torra (2024); Holohan et al. (2017), and similar works, which combine k-anonymity and DP into a unified framework.

To achieve (ϵ, δ)-DP as defined in Equation 3, each original embedding $\mathbf{x}$ is first clipped to a norm bound B, so that $||\mathbf{x}||_2 \leq B$. Then, cluster centroids are computed from these clipped embeddings. This gives an upper bound to the sensitivity of the centroid computation function. Following the definition in Equation 5, the sensitivity Δf_j in this case can be interpreted as the maximum change in the centroid of a cluster j resulting from the inclusion or exclusion of a single embedding $\mathbf{x}$ in j. Bounding this sensitivity ensures that no individual data point can significantly influence the output, thereby reducing re-identification risk.

The next step involves adding Gaussian noise $\mathbf{z}_j \sim \mathcal{N}(\mathbf{0}, \sigma_j^2 \mathbf{I})$ to the centroid of each cluster j. The standard deviation σ_j is calculated according to the work of Dwork and Roth (2013) as:

$$\sigma_j = \frac{\Delta f_j \cdot \sqrt{2 \ln(1.25/\delta)}}{\epsilon}. \tag{6}$$

Leveraging Global Knowledge Inspired by Garg and Torra (2024) and Holohan et al. (2017), we refer to the resulting anonymized dataset as a (k, ϵ, δ)-anonymized dataset $\mathcal{S}'_m$, where all embeddings in each cluster are represented by its DP-noised centroid. The server collects all $\mathcal{S}'_m$ into a global pool $\hat{\mathcal{S}}^{\mathrm{g}} = \bigcup_{m=1}^{M} \mathcal{S}'_m$. Each client receives this pooled set, excluding its own share. This ensures each client has access to a diverse, privacy-preserving global dataset for downstream tasks. Algorithm 2 outlines this DP-kSame sharing process.

In summary, DP-kSame offers several advantages over existing prototype-based FL: (i) it preserves each client's unique information by sharing anonymized but not aggregated centroids (as opposed to averaged class prototypes), (ii) it provides formal privacy guarantees via k-anonymity and DP, and (iii) it leverages powerful FMs and non-parametric clustering to produce compact, semantically rich representations for global data sharing without requiring local training of an embedding model.

Algorithm 2 `DP-kSame`

Each client $m \in [M]$ creates a (k, ϵ, δ)-anonymized dataset $\mathcal{S}'_m$ from its local embedding set $\mathcal{S}_m$ using DP-noised cluster centroids.

Input: Embedding dataset $\mathcal{S}_m = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n_m}$; anonymity parameter k; DP parameters ϵ, δ and norm bound B
Output: Anonymized dataset $\mathcal{S}'_m$ of same size n_m

// Step 1: Feature Clipping
for each $\mathbf{x}_i \in \mathcal{S}_m$ **do**
    $\bar{\mathbf{x}}_i \leftarrow \mathbf{x}_i / \max(1, \frac{||\mathbf{x}_i||_2}{B})$ ▷ clip ℓ_2 norm of feature vectors to B
end for

// Step 2: Unsupervised Clustering
$J \leftarrow \max(1, \lfloor \frac{n_m}{k} \rfloor)$ ▷ get number of clusters of size k
Form J clusters from $\{\bar{\mathbf{x}}_i\}$ and get centroids $\{C_m^{(j)}\}_{j=1}^{J}$
$\forall j \in [J] : \hat{y}^{(j)} \leftarrow \arg\max_{c \in \mathcal{Y}} \sum_{\bar{\mathbf{x}}_i \in \text{cluster } j} \mathbb{1}_{\{y_i = c\}}$ ▷ assign label to centroid $C_m^{(j)}$

// Step 3: Gaussian DP Noise to Centroids
for each cluster $j = 1, \ldots, J$ **do**
    $\Delta_j \leftarrow \frac{2B}{k_j}$ ▷ compute sensitivity of cluster mean with k_j points
    $\sigma_j \leftarrow \frac{\Delta_j \cdot \sqrt{2\ln(1.25/\delta)}}{\epsilon}$ ▷ compute standard deviation for Gaussian distribution
    $\mathbf{z}_j \sim \mathcal{N}(\mathbf{0}, \sigma_j^2 \mathbf{I})$ ▷ sample noise from Gaussian distribution
    $\tilde{C}_m^{(j)} \leftarrow C_m^{(j)} + \mathbf{z}_j$ ▷ add noise to centroids
end for

// Step 4: k-Same Anonymization
for each $\bar{\mathbf{x}}_i \in$ cluster j **do**
    $\mathbf{x}'_i \leftarrow \tilde{C}_m^{(j)}$ ▷ replace each real embedding in a cluster with its centroid
end for

return $\mathcal{S}'_m = \{(\mathbf{x}'_i, \hat{y}^{(j)}) : \bar{\mathbf{x}}_i \in \text{cluster } j\}_{i=1}^{n_m}$

3.2.2 Synthetic Data Generation via DP-CVAE

While DP-kSame provides formal privacy, cluster centroids are inherently coarse approximations of the original data and may fail to capture fine-grained structure.

Anonymization methods relying on k-anonymity also tend to reduce data diversity, which can negatively impact model robustness (Di Salvo et al., 2024; Larson et al., 2020; Yu et al., 2022). Although reducing k improves fidelity, it increases the risk of re-identification (Abadi et al., 2016; Aggarwal, 2005; Narayanan and Shmatikov, 2008). Furthermore, the direct addition of DP noise to centroids can distort data distributions and harm downstream performance. These limitations motivate a more expressive yet still privacy-preserving alternative.

To retain richer information, we propose sharing synthetic embeddings produced by a federated, differentially private generative model. FM embeddings are semantically rich and lower-dimensional than raw images, which reduces training complexity and improves the quality of generated samples. A recent study (Di Salvo et al., 2024) has shown that training on synthetic embeddings can preserve classification accuracy compared to training on raw images while enhancing data privacy.

Among generative models, GANs and diffusion models are often computationally intensive and thus impractical for resource-constrained federated environments. GANs, in particular, are prone to training instabilities and mode collapse, leading to reduced sample diversity (Fonseca and Bação, 2023; Koetzier et al., 2024; Hayaeian Shirvan et al., 2025). In contrast, VAEs and CVAEs are more stable and lightweight. While they may produce slightly blurrier outputs, CVAEs avoid the complexity and instability of adversarial training and offer better control and reproducibility. Di Salvo et al. (2024) also shows that CVAE-generated embeddings can exceed k-Same datasets in diversity and robustness.

CVAE We adopt a CVAE architecture inspired by Di Salvo et al. (2024), with symmetric encoder and decoder networks, each comprising three linear layers. Extending the standard VAE framework (Kingma and Welling, 2014), the CVAE models the conditional generation of an embedding $\mathbf{x}$ given a class label y via a latent variable $\mathbf{z}$. Specifically, the encoder, represented by $q_\phi(\mathbf{z} \mid \mathbf{x}, y)$, approximates the true posterior over $\mathbf{z}$ given $\mathbf{x}$ and y, outputting the parameters (*i.e.,* the mean and standard deviation) of a Gaussian distribution. The decoder $p_\theta(\mathbf{x} \mid \mathbf{z}, y)$ reconstructs $\mathbf{x}$ from $\mathbf{z}$ and y.

The model samples $\mathbf{z}$ from the encoder distribution using the reparameterization trick (Kingma and Welling, 2014) and generates a reconstruction $\mathbf{x}^*$. Training maximizes a variational lower bound on the conditional log-likelihood $\log p_\theta(\mathbf{x} \mid y)$, which decomposes into: (i) a reconstruction loss, for which we use the Mean Squared Error (MSE) between $\mathbf{x}$ and $\mathbf{x}^*$, and (ii) a regularization term, which is the Kullback–Leibler (KL) divergence between the learned latent distribution $q_\phi(\mathbf{z} \mid \mathbf{x}, y)$ and the prior $p_\theta(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. A hyperparameter β controls the trade-off between reconstruction fidelity and alignment of the learned latent distribution with the prior (indicating better sampling and generalization). The resulting CVAE loss function for a single embedding–label pair $(\mathbf{x}, y)$ is:

$$\mathcal{L}_{\text{CVAE}}(\mathbf{x}, y; \theta, \phi) = \text{MSE}(\mathbf{x}, \mathbf{x}^*) - \beta \cdot D_{\text{KL}}\left(q_\phi(\mathbf{z} \mid \mathbf{x}, y) \| p_\theta(\mathbf{z})\right), \tag{7}$$

where $\mathbf{x}^* \sim p_\theta(\mathbf{x} \mid \mathbf{z}, y)$, and $\mathbf{z} \sim q_\phi(\mathbf{z} \mid \mathbf{x}, y)$. The loss encourages the decoder to accurately reconstruct embeddings of specific classes while regularizing the latent space to remain close to the standard normal prior for effective generative sampling.

To generalize well across clients while preserving privacy, we train the CVAE in a federated manner. Each client m maintains a local encoder with parameters ϕ_m and a shared decoder with parameters θ. At each round, clients train on their local dataset $\mathcal{S}_m$, updating both CVAE components, but share only the decoder parameters with the server. In FedAvg style (see Algorithm 1), the shared decoder weights are averaged to update the global decoder $\mathcal{D}^{\mathrm{g}}$, which is redistributed to clients. Over rounds, $\mathcal{D}^{\mathrm{g}}$ captures global (intra-class) variations, while the encoder stays personalized. For comparison, we also experiment with a CGAN model where the generator is federated and the discriminator is local.

Enhancing Privacy with DP In training federated generative models, DP is a well-established method for ensuring provable privacy guarantees through controlled injection of random noise (Kaabachi et al., 2025). To achieve (ϵ, δ)-DP during our CVAE training, we integrate DP-SGD algorithm Abadi et al. (2016), illustrated in Figure 5. Each client clips per-sample gradients of the decoder to a norm bound B, aggregates them, and adds Gaussian noise $\mathcal{N}(\mathbf{0}, \sigma^2 B^2 \mathbf{I})$ before updating the global decoder. Any outputs of the resulting decoder are guaranteed to protect an individual's data used in training according to the chosen privacy budget (ϵ, δ). This method mirrors the DP-noise mechanism used in DP-kSame but applies it to gradients instead of feature embeddings (see Table 1 for a full comparison). In both approaches, privacy-preserving operations are performed locally by each client without exposing individual privacy details to the server.

Table 1: Comparison of DP mechanisms in our federated data sharing methods

	DP-kSame	DP-CVAE/-CGAN
Clipping target	Per-sample embeddings	Per-sample gradients
Noise target	Cluster centroids	Aggregated (average) gradients
Sensitivity based on	Norm of embeddings	Norm of per-sample gradients

Leveraging Global Knowledge Once the federated training of the global decoder $\mathcal{D}^{\mathrm{g}}$ is completed, each client uses it to create a synthetic dataset that approximates a globally representative, privacy-preserving feature distribution. Given a target synthetic dataset size N_m, each client m constructs its synthetic dataset $\hat{\mathcal{S}}_m$ as follows:

$$\hat{\mathcal{S}}_m = \{(\hat{\mathbf{x}}_i, \hat{y}_i)\}_{i=1}^{N_m}, \qquad \hat{\mathbf{x}}_i = \mathcal{D}^{\mathrm{g}}(\mathbf{z}_i \mid \hat{y}_i) \tag{8}$$

where $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is sampled from a standard normal Gaussian distribution, and $\hat{y}_i$ is sampled from a selected class distribution K. This synthetic dataset can then

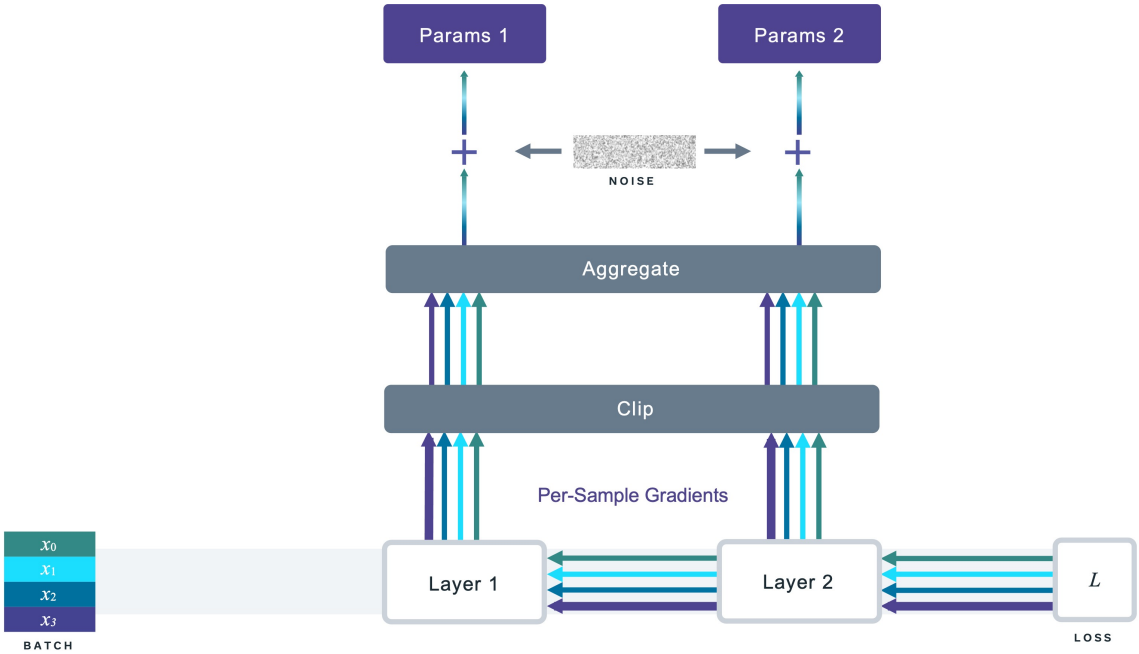

Figure 5: Visualization of the DP-SGD algorithm. Solid-colored lines represent per-sample gradients (with width indicating norm), and multicolored lines represent the aggregated, noised gradients. This figure is reproduced from Yousefpour et al. (2022).

be used for downstream tasks. Algorithm 3 details the training and generation processes in our DP-CVAE approach.

3.3 Downstream Classification

All downstream tasks (*e.g.,* classification) are performed in the embedding space of a pre-trained FM. This enables clients to achieve strong performance even with simple classifiers such as linear probes or non-parametric methods like k-NN. Importantly, in this setup, downstream training is decoupled from the federated aggregation process, offering several advantages as discussed in Section 2.1. For example, clients gain autonomy (as no parameter synchronization is required) and can rapidly adapt to new classes or tasks. Prior work such as Doerrich et al. (2024) demonstrated that combining k-NN with vision FMs yields adaptable, privacy-preserving classifiers that outperform standard deep networks in certain settings.

In both DP-kSame and DP-CVAE, each client m retains two datasets: its private local set $\mathcal{S}_m$ and a global set $\hat{\mathcal{S}}_m$ (anonymized or synthetic). This setup enables clients to tailor models to their specific downstream needs. For instance, clients can utilize the same synthetic data to train classifiers with different label granularities, model data distributions for anomaly detection, or address out-of-distribution detection, depending on the client's application.

Algorithm 3 `DP-CVAE`

Clients train a CVAE in a differentially private manner, and synchronize only the decoder via FedAvg. The final global decoder is used to generate synthetic data locally.

Input: M participating clients with local embedding datasets $\mathcal{S}_m = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n_m}$; number of communication rounds T; number of local training epochs E; learning rate η; DP noise scale σ (determined based on input (ϵ, δ)), clipping norm B; batch size n_b; target synthetic size N_m; class distribution K

Output: Global decoder $\mathcal{D}^{\mathrm{g}}$, synthetic dataset $\hat{\mathcal{S}}_m$ for each client m

// Step 1: Federated Training of DP-CVAE

Server executes:

Initialize weights $\theta^{(0)}$ for global decoder $\mathcal{D}^{\mathrm{g}}$

for each round $t = 1, \ldots, T$ **do**

  for each client $m = 1, \ldots, M$ **in parallel do**

    $\theta_m^{(t)} \leftarrow \textsc{ClientUpdate}(m, \theta^{(t)})$

  end for

  $s_t \leftarrow \sum_{m=1}^{M} n_m^{\text{train}}$

  $\theta^{(t+1)} \leftarrow \sum_{m=1}^{M} \frac{n_m^{\text{train}}}{s_t} \theta_m^{(t)}$ ▷ FedAvg on CVAE decoder only

end for

function $\textsc{ClientUpdate}(m, \theta)$

  Initialize or load encoder ϕ_m; set decoder $\theta_m \leftarrow \theta$

  for each local epoch $e = 1, \ldots, E$ **do**

    for each mini-batch $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n_b} \subset \mathcal{S}_m$ **do**

      For each $(\mathbf{x}_i, y_i)$, compute $\mathbf{g}_i \leftarrow \nabla_{\theta_m, \phi_m} \mathcal{L}_{\text{CVAE}}(\mathbf{x}_i, y_i; \theta_m, \phi_m)$

      $\bar{\mathbf{g}}_i \leftarrow \mathbf{g}_i / \max(1, \frac{\|\mathbf{g}_i\|_2}{B})$ ▷ clip per-sample gradients

      $\tilde{\mathbf{g}} \leftarrow \frac{1}{n_b} \sum_i \bar{\mathbf{g}}_i + \mathcal{N}(\mathbf{0}, \sigma^2 B^2 \mathbf{I})$ ▷ add noise to aggregated gradient

      $(\theta_m, \phi_m) \leftarrow (\theta_m, \phi_m) - \eta \cdot \tilde{\mathbf{g}}$ ▷ update CVAE with DP gradient

    end for

  end for

  return θ_m to server

end function

// Step 2: Local Synthetic Data Generation

for each client $m = 1, \ldots, M$ **do**

  $\hat{\mathcal{S}}_m \leftarrow \emptyset$

  for $i = 1, \ldots, N_m$ **do**

    $\hat{y}_i \sim K$ ▷ sample label from selected class distribution

    $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ ▷ sample latent from standard normal distribution

    $\hat{\mathbf{x}}_i \leftarrow \mathcal{D}^{\mathrm{g}}(\mathbf{z}_i \mid \hat{y}_i)$ ▷ generate embedding

    $\hat{\mathcal{S}}_m \leftarrow \hat{\mathcal{S}}_m \cup \{(\hat{\mathbf{x}}_i, \hat{y}_i)\}$

  end for

end for

return $(\hat{\mathcal{S}}_m)_{m=1}^{M}$

We focus on image classification as the primary downstream task for a standardized evaluation. The availability of both local and global datasets supports *personalized* FL, where each client trains a model adapted to its local distribution while leveraging knowledge from the broader federation. To formalize our objective, we begin with the optimization problem of *parametric* personalized FL, which can be defined as follows:

$$\forall m \in [M], \quad \theta_m^* = \arg\min_{\theta_m} \mathcal{L}_m(h_m^{\theta_m}), \tag{9}$$

where $\mathcal{L}_m(h_m^{\theta_m}) = \mathbb{E}_{(\mathbf{x},y)\sim P_m}\left[\ell\left(h_m\left(\mathbf{x};\theta_m\right),y\right)\right]$ is the error of client m's model h_m with parameters θ_m in predicting the true label $y \in \mathcal{Y}_m$ given the input $x \in \mathcal{X}_m$, and P_m is the distribution over $\mathcal{X}_m \times \mathcal{Y}_m$ (Chen and Chao, 2022; Morafah et al., 2024). The objective of parametric personalized FL is therefore to train models to perform well on each client's distinctive data distribution. Marfoq et al. (2022) generalizes this objective to *non-parametric* models:

$$\forall m \in [M], \quad h_m^* = \arg\min_{h_m \in \mathcal{H}} \mathcal{L}_m(h_m), \tag{10}$$

where $h_m \in \mathcal{H}$ is a model (hypothesis) fit to each client's data distribution, mapping each input $x \in \mathcal{X}_m$ to a probability distribution over $\mathcal{Y}$; the true risk of a model h_m under data distribution P_m is measured by $\mathcal{L}_m(h_m) = \mathbb{E}_{(\mathbf{x},y)\sim P_m}\left[\ell\left(h_m\left(\mathbf{x}\right),y\right)\right]$. h_m can therefore be interpreted as an estimate of the conditional probability distribution $P_m(y \mid \mathbf{x})$, and non-parametric personalized FL directly optimizes over the hypothesis space $\mathcal{H}$ without explicit parameters.

In contrast to many existing personalized FL methods (cf. Section 2.2), our approach leverages shared feature representations from a FM and focuses solely on downstream training without additional representation learning. For each client m, we train one classifier on the local data $\mathcal{S}_m$ and another on the global (anonymized or synthetic) data $\hat{\mathcal{S}}_m$. Predictions are then obtained by interpolating the outputs of these two models using a tunable parameter $\lambda_m \in [0,1]$, optimized based on the validation set. Given a test sample $\mathbf{x}_{\text{test}}$, the interpolated class distribution is:

$$h_{m,\lambda_m}(y \mid \mathbf{x}_{\text{test}}) := \lambda_m P_{\mathcal{S}_m}(y \mid \mathbf{x}_{\text{test}}) + (1-\lambda_m) P_{\hat{\mathcal{S}}_m}(y \mid \mathbf{x}_{\text{test}}) \tag{11}$$

where $P_{\mathcal{S}_m}$ and $P_{\hat{\mathcal{S}}_m}$ denote the predictive distributions from the local and global models, respectively. The final predicted class is given by:

$$\hat{y}_{\text{test}} = \arg\max_{c \in \mathcal{Y}} P_{m,\lambda_m}(y \mid \mathbf{x}_{\text{test}}) \tag{12}$$

As $\lambda_m \to 1$, the model emphasizes client-specific knowledge. Conversely, as $\lambda_m \to 0$, it relies more on global information aggregated across clients. Thus, λ_m offers a tunable balance between personalization and generalization.

4 Experiments

4.1 General Setup

4.1.1 Datasets

We evaluate DP-kSame and DP-CVAE on a suite of medical image datasets to assess their performance in privacy-sensitive and realistic federated learning scenarios. Specifically, we use multiple 2D multi-class classification datasets from MedMNISTv2 (Yang et al., 2023), along with a subset of 4,600 images from Camelyon17-Wilds (Koh et al., 2021), as utilized in Chen et al. (2023b). Camelyon17 is a binary histopathological image dataset drawn from five hospitals in the Netherlands, naturally reflecting a federated setting. The selected datasets vary in different aspects like imaging modality, resolution, and class balance, posing distinct challenges in learning and privacy protection. They provide a comprehensive benchmark for evaluating privacy-utility trade-offs and potential robustness of our methods in diverse clinical imaging tasks. Table 2 gives an overview of the datasets used.

Table 2: Overview of benchmark medical datasets used in experiments

Dataset	Data Modality	# Classes	# Samples
Camelyon17	Histopathology	2	4,600
OrganSMNIST	Abdominal CT	11	25,211
BloodMNIST	Blood Cell Microscopy	8	17,092
DermaMNIST	Dermatoscopy	7	10,015
PneumoniaMNIST	Chest X-Ray	2	5,856

For feature extraction, we opt for the DINOv2-Base model as our primary backbone due to its state-of-the-art performance in learning general-purpose visual features. Using this backbone, we obtain 768-dimensional embeddings from our image datasets. For Camelyon17 dataset, each hospital is treated as a distinct client. In the MedMNIST datasets, we simulate heterogeneity by partitioning each dataset among M clients using a Dirichlet distribution, following Marfoq et al. (2022). For each class label y, we draw a vector $p_y = [p_{y,1}, \ldots, p_{y,M}]$ from a symmetric Dirichlet distribution with parameter α and assign a fraction $p_{y,m}$ of samples with label y to client $m \in [M]$. Smaller values of α result in more skewed (non-IID) distributions across clients. Each client's data is split into training (70%), validation (10%), and test (20%) sets.

4.1.2 Implementation Details

DP-kSame Each client anonymizes its embedding dataset using the k-Same algorithm with $k = 10$, grouping roughly 10 samples per cluster. The clustering (via k-means) is performed in an unsupervised manner to obscure class identity. Cluster centroids are used to replace the embeddings in the corresponding clusters, and their labels are assigned via majority voting. To satisfy (ϵ, δ)-DP with $(1.0, 10^{-4})$, Gaussian noise is added to each centroid after clipping real embeddings with ℓ_2-norm bound 1.5. This choice, with $\epsilon \leq 1$ and $\delta \ll 1/n$, where n is the average number of training samples per client in our experiments, provides meaningful privacy guarantees while preserving utility (Nasr et al., 2021; Lange et al., 2022).

DP-CVAE Clients train class-conditional CVAEs on their local embeddings. The CVAE has three linear layers in both encoder and decoder, with a latent dimension of 100. Class conditioning is implemented via concatenation of one-hot encoded labels to the inputs at both ends. The model is trained for 50 communication rounds with 5 local epochs per round, using Adam optimization with learning rate of 10^{-3}. During local training, we apply DP-SGD via the Opacus library (Yousefpour et al., 2022) with $(\epsilon, \delta) = (1.0, 10^{-4})$ and a clipping norm of 1.5, similar to the DP configuration used in DP-kSame. Clients share only their decoder weights, and the server aggregates them via FedAvg into a global decoder $\mathcal{D}^{\mathrm{g}}$.

After training, each client generates synthetic embeddings by sampling latent vectors from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and decoding them with $\mathcal{D}^{\mathrm{g}}$. Instead of reproducing the original class distribution, we exploit the flexibility of this generative approach to explore alternative class-balancing strategies during generation – such as uniform or inverse-frequency sampling – while keeping the total number of generated samples equal to the original dataset size for a fair comparison with DP-kSame. Since the trained CVAE can produce an arbitrary number of samples for any class, it enables the construction of synthetic datasets tailored to each client's needs, which can then be combined with the client's real embeddings to train downstream models. This generative flexibility is a key advantage of our DP-CVAE approach. For all experiments reported below, we adopt inverse class frequency weighting (*i.e.,* oversampling minority classes), which consistently yields stronger classification performance. By generating diverse variations of input samples, the CVAE increases intra-class diversity, particularly benefiting underrepresented classes.

Downstream Classification We evaluate both linear classifiers and non-parametric k-NN models. Linear models are trained using the Adam optimizer with a learning rate of 10^{-3} for 100 epochs. For k-NN, we use the FAISS library (Douze et al., 2024) for efficient search with Euclidean distance metric. The number of nearest neighbors is set to $k = 3$.

Like linear classifiers, k-NN is applied separately on local and global data to compute a local probability $P_{\mathcal{S}_m}(y \mid \mathbf{x}_{\text{test}})$ and a global probability$P_{\hat{\mathcal{S}}_m}(y \mid \mathbf{x}_{\text{test}})$. While the outputs of linear classifiers can be interpreted as softmax logits, the output of k-NN is a probability vector over classes, where each entry reflects the weighted influence of the k nearest neighbors for a given class. We compute these weights using a

Gaussian kernel applied to the neighbors' distances to the test embedding, which transforms distances into smoothly decaying weights. This approach helps mitigate sensitivity to variations in distance and reduces the influence of outliers. For each set of k nearest neighbors retrieved from the datastore (local or global), we compute:

$$P(y = c \mid \mathbf{x}_{\text{test}}) = \frac{\sum_{i=1}^{k} \mathbb{1}_{\{y_i=c\}} \cdot \exp\left\{-d\left(\mathbf{x}_{\text{test}}, \mathbf{x}_i\right)\right\}}{\sum_{i=1}^{k} \exp\left\{-d\left(\mathbf{x}_{\text{test}}, \mathbf{x}_i\right)\right\}} \tag{13}$$

where $d\left(\mathbf{x}_{\text{test}}, \mathbf{x}_i\right)$ denotes the Euclidean distance between the test embedding and neighbor i. For final predictions, interpolation weights $\lambda_m \in 0.0, 0.1, \ldots, 1.0$ are selected for each client via validation performance to combine the global and local model outputs, as described in Equation 11.

4.2 Downstream Classification Performance

4.2.1 Benchmarks

We compare DP-kSame and DP-CVAE against both standard and personalized FL baselines. The standard methods include FedAvg (McMahan et al., 2017) and its enhanced version, FedProx (Li et al., 2020), which addresses local data heterogeneity by adding a regularization term that penalizes large deviations in client updates from the global model. Personalized FL baselines include kNN-Per (Marfoq et al., 2022) and FedProto (Tan et al., 2022). Similar to our methods, these approaches aim to obtain personalized models for participating clients and shift towards decoupling global and local training, as analyzed in Section 2.2.

For kNN-Per, we implement an adapted version where each client uses the pre-trained FM instead of the global model to obtain embeddings. The outputs of the global model are now used solely for the downstream task by interpolating them with the local k-NN classifier to produce the final prediction. kNN-Per remains our *semi-parametric* competitor, as the global model is still trained through FedAvg for the downstream task. The other personalized baseline, FedProto, closely resembles our methods, particularly DP-kSame, in its prototype-based data sharing approach, as described in Section 3.2.1. We adapt FedProto by computing fixed class prototypes using frozen FM feature representations, rather than iteratively updating them via client-side model training. This change reflects our use of a single-layer linear classifier on top of fixed embeddings, in contrast to training full deep networks. As a result, downstream training becomes entirely local and the overall pipeline becomes non-parametric. This effectively makes FedProto our *non-parametric* baseline, where local linear classifiers are trained on frozen FM embeddings and use static class prototypes for regularization.

We also compare against DP-CGAN (Sun et al., 2023; Torkzadehmahani et al., 2019), a generative baseline trained with DP-SGD. Here, a global generator is learned in a federated manner by aggregating generator weights across clients and

Table 3: Mean ± standard deviation of test accuracy (ACC) and balanced accuracy (BACC) across clients, averaged over three seed runs. We highlight in **bold** the top two results in each setting. Rows containing methods that incorporate privacy-enhancing mechanisms are shaded in gray. For these, both k-NN and linear classifiers are evaluated.

	OrganSMNIST (IID)		OrganSMNIST ($\alpha = 0.3$)		Camelyon17
	ACC	BACC	ACC	BACC	ACC
FedAvg	72.29±0.51	65.69±0.73	64.57±7.35	**59.28±4.48**	88.77±3.42
FedProx	72.24±1.04	65.64±1.04	64.44±9.92	59.20±4.18	88.70±3.50
kNN-Per	75.22±0.95	68.93±1.24	87.21±1.39	**60.39±3.12**	92.21±2.93
FedProto	76.87±1.06	**70.98±1.22**	88.92±2.27	57.00±2.55	94.09±1.22
DP-kSame (Ours)					
+ k-NN	71.48±1.81	65.90±2.40	86.46±2.56	56.15±2.99	92.07±1.61
+ Linear	76.41±1.16	70.54±1.27	**89.18±1.23**	58.39±2.12	92.21±1.20
DP-CGAN					
+ k-NN	71.64±1.45	66.01±1.85	86.43±1.85	56.27±2.80	92.03±0.93
+ Linear	**76.90±1.14**	70.90±1.87	88.93±1.28	57.53±2.61	**94.09±0.94**
DP-CVAE (Ours)					
+ k-NN	71.66±1.32	66.14±1.84	86.53±1.22	56.27±2.56	91.99±1.29
+ Linear	**76.96±0.99**	**71.14±1.27**	**89.03±0.74**	57.58±3.51	**94.57±0.91**

used for producing synthetic embeddings. Like DP-kSame and DP-CVAE, we evaluate DP-CGAN's downstream performance with both k-NN and linear classifiers.

All classifiers except kNN-Per's local k-NN classifiers are implemented as single-layer linear models. Similar to our DP-CVAE, all federated models are trained for 50 rounds with 5 local epochs using SGD optimization with a learning rate of 10^{-3}. Local linear classifiers in FedProto are trained for 100 epochs using the Adam optimizer with a learning rate of 10^{-3}, as in DP-kSame and DP-CVAE.

4.2.2 Average Performance

Table 3 reports the overall mean and standard deviation of classification accuracy and balanced accuracy across clients and three random seeds. Following Li et al.

(2024b), we use OrganSMNIST – an abdominal CT dataset (Sagittal view) (Xu et al., 2019), utilizing the splits from Yang et al. (2023) with $25,211$ images across 11 classes. We distribute the data among 10 clients under both highly non-IID ($\text{Dir}(\alpha = 0.3)$) and IID conditions. In addition, we test on Camelyon17, adopting its natural partitioning across five hospitals as five clients. Since Camelyon17 is roughly class-balanced, we report only the accuracy.

For all interpolation-based methods (kNN-Per, DP-kSame, DP-CGAN, DP-CVAE) and settings reported in Table 3, the optimal λ_m (averaged across clients over three seed runs) is around 0.7, indicating that a 70% weight on the local model yields the best results. This observation enables a fair comparison across methods, as it ensures a similar balance between local and global contributions to final predictions.

Overall, personalized FL methods improve substantially upon FedAvg and FedProx, demonstrating the benefits of incorporating local predictions for personalization. Among these, federated data sharing schemes outperform end-to-end federated training approaches, with DP-CVAE achieving the best performance in most cases, closely followed by DP-CGAN, DP-kSame and FedProto. This supports our hypothesis that sharing anonymized or synthetic embeddings, even under DP constraints, can be more effective than exchanging model parameters. The success of these methods despite privacy-enhancing mechanisms highlights the benefits of decoupling global learning from local downstream tasks. As argued by Marfoq et al. (2022), this decoupling avoids conflicts between global and local objectives.

Among privacy-preserving methods, DP-CVAE performs best overall, especially in data-scarce settings like Camelyon17, where each client has fewer than 500 training samples on average. Its notable performance gap over DP-kSame and DP-CGAN on Camelyon17 suggests that CVAEs are more data-efficient than the other anonymization mechanisms. The success of DP-CVAE highlights the effectiveness of our generative modeling approach in preserving utility while providing strong privacy protection.

In highly heterogeneous settings like OrganSMNIST with $\alpha = 0.3$, our DP-kSame and DP-CVAE methods achieve slightly lower balanced accuracy than FedAvg, FedProx, and kNN-Per, yet surpass them in overall accuracy. This discrepancy arises from the nature of the data used for downstream training: while traditional federated classifiers and kNN-Per rely entirely on clients' true local data distributions, DP-kSame and DP-CVAE train models on data shaped by global aggregation – either globally collected anonymized centroids or synthetic samples from federated decoders. The resulting global models, interpolated with local models at inference, generalize well across clients. However, in extreme non-IID scenarios, the global classifier may underrepresent minority classes specific to a client. As a result, its contribution can dampen the balanced accuracy of the interpolated model. Despite this, the accuracy gains show that modeling shared structure through global data remains beneficial. Furthermore, it is important to note that the benchmark methods are not differentially private, whereas DP-kSame and DP-CVAE enforce strong privacy guarantees, which may affect performance. The additive DP noise can dis-

tort the representation space, leading to less precise decision boundaries, especially for underrepresented classes.

Although our discussion mainly focuses on linear models due to their overall better performance, it is noteworthy that a simple k-NN classifier performs nearly on par with a linear model trained for 100 epochs. The gap with linear classifiers is especially small in highly non-IID conditions like OrganSMNIST ($\alpha = 0.3$). confirming the strength of rich FM embeddings (Caron et al., 2021; Oquab et al., 2024). This aligns with earlier evaluations of DINO features, suggesting that rich FM embeddings allow a basic k-NN model to achieve competitive results without additional training (Caron et al., 2021; Oquab et al., 2024). Moreover, as noted by Marfoq et al. (2022), k-NN classifiers offer fast adaptability by updating the local datastore without requiring a full model retraining cycle.

Lastly, across methods, the tuned interpolation weight λ_m yielding the best (balanced) accuracy consistently lies between 0.4 and 0.7. This confirms that combining global and local models provides a consistent performance boost and underscores the utility of the shared anonymized or synthetic data in our methods.

4.2.3 Effect of Data Heterogeneity

Figure 6 shows how balanced accuracy changes with data heterogeneity (defined by Dirichlet parameter α) for four MedMNIST datasets. We exclude Camelyon17 due to its fixed client structure. For each α, we simulate 10 clients and report the average balanced accuracy across clients. To compare the best possible performance of our privacy-preserving methods against the benchmarks, we use linear models instead of k-NN for downstream classification in these methods.

As expected, performance generally improves as data become more IID (larger α) and converge as $\alpha \to \infty$ (near IID). For OrganSMNIST and BloodMNIST, FedAvg and kNN-Per seem to address high heterogeneity ($\alpha \leq 0.5$) more effectively, but DP-kSame and DP-CVAE (along with FedProto) gain a clear advantage in more balanced settings ($\alpha \geq 1$). This suggests that the shared (synthetic or anonymized) data improves generalization compared to training solely on local data. This result aligns with our observations in the previous experiment that the current design of our DP-kSame and DP-CVAE works best in less heterogeneous settings, where they can fully leverage globally aggregated knowledge and shared structure for downstream prediction.

Notably, DP-CVAE consistently outperforms DP-kSame, particularly under high heterogeneity or in limited-data scenarios like DermaMNIST and PneumoniaMNIST. This further underscores our earlier observation from the classification results on Camelyon17 about the robustness of our generative method in challenging data conditions. Importantly, DP-CVAE's flexibility in sample generation allows future extensions to more advanced class-rebalancing strategies (Liu et al., 2025), unlike the fixed anonymization scheme in DP-kSame.

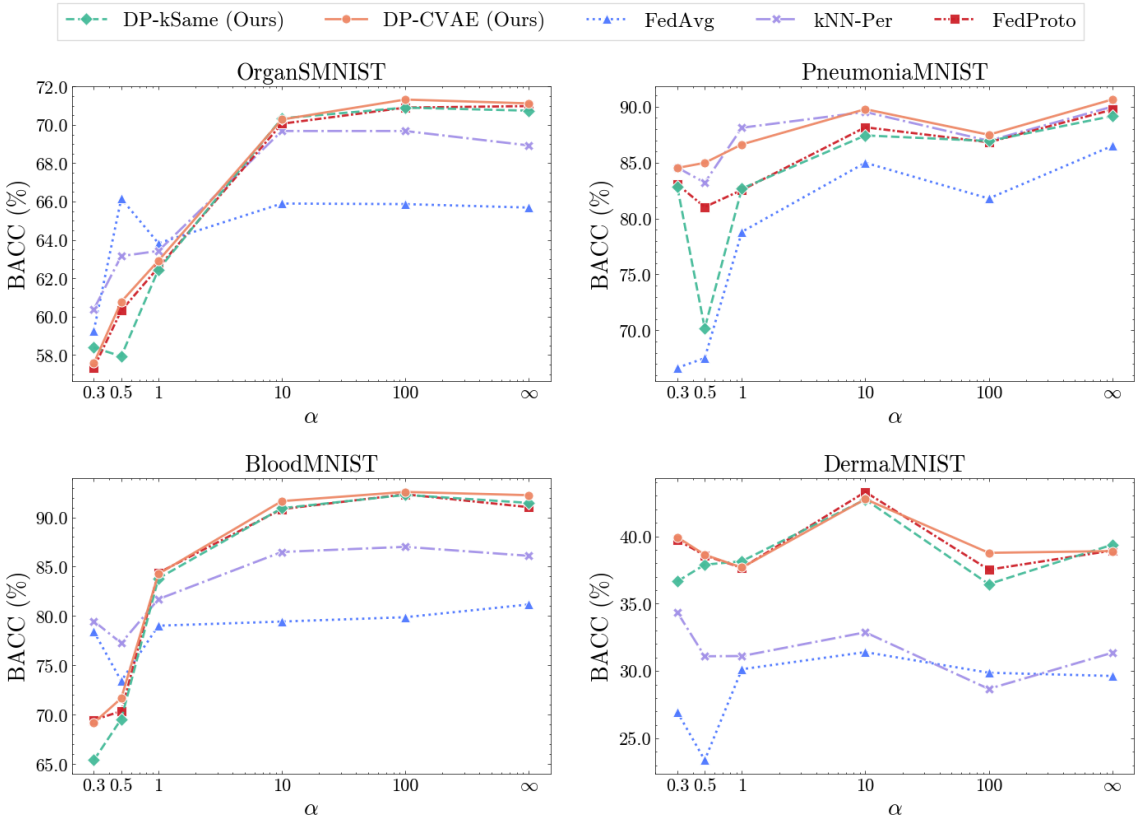

Figure 6: Balanced accuracy (BACC) (averaged across 10 clients over three seed runs) versus Dirichlet parameter α. Smaller α implies higher heterogeneity; as $\alpha \to \infty$, local distributions approach the IID setting.

4.2.4 Generalizability across Backbones and Robustness to Data Scarcity

In Figure 7, we demonstrate the robustness of our methods across different backbones (DINO-, DINOv2-, ViT-Small) and varying client counts. We use OrganSMNIST under IID partitioning.

The results show that both methods generalize well across different feature extractors, with a modest drop in balanced accuracy from DINO to DINOv2 to ViT. As the number of clients increases, resulting in smaller local datasets, performance declines only marginally. Compared to the upper-bound performance achievable through centralized learning, our methods remain competitive while ensuring privacy, even with limited data. Overall, the experiment demonstrates that our methods are robust across diverse feature representations and dataset sizes, with DP-CVAE consistently outperforming DP-kSame in most settings. This further suggests a favorable privacy-utility trade-off: our methods achieve competitive predictive performance while providing strong privacy protection, generalizing nearly as well as models trained on real embeddings.

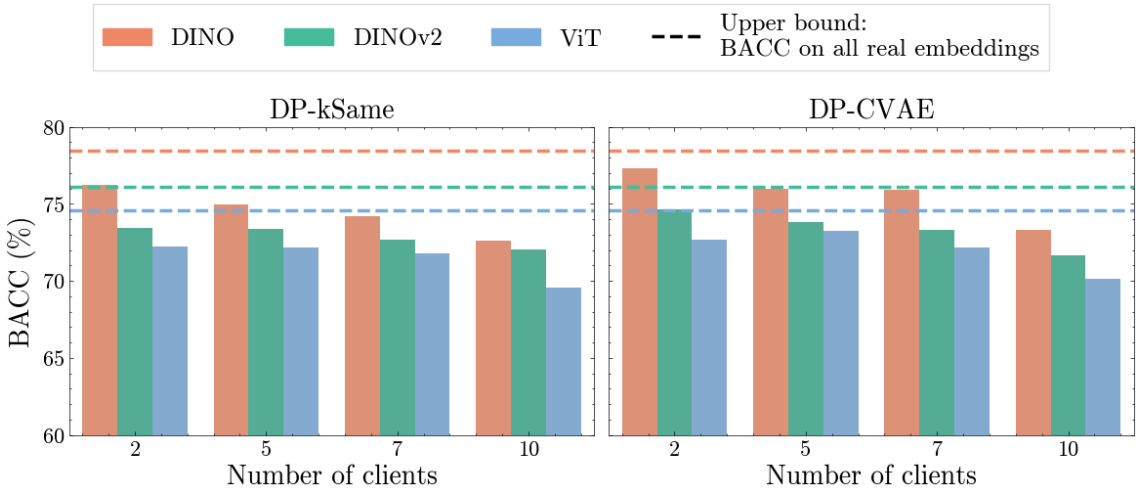

Figure 7: Generalizability of our methods across different backbones and increasing numbers of clients (*i.e.*, fewer samples per client), utilizing OrganSMNIST (IID). The dashed lines represent the BACC obtained when training a linear classifier on real image embeddings with the original train–validation–test split for each backbone.

4.3 Privacy-Utility Trade-off

4.3.1 Data Utility across Differential Privacy Budgets

To evaluate the trade-off between privacy and utility in our proposed methods, we examine how the fidelity of the shared data deteriorates as we tighten the DP budget (ϵ), which is used in data anonymization (DP-kSame) or privacy-preserving generative modeling (DP-CVAE). We consider Camelyon17 and OrganSMNIST datasets without differentiating between IID and non-IID settings due to similar results. Specifically, we vary ϵ (with a fixed $\delta = 10^{-4}$) and measure the 2-Wasserstein distance (Panaretos and Zemel, 2019) between a client's real embedding distribution and the distribution of its shared data.

The p-Wasserstein distance, also known as the Earth Mover's Distance, quantifies the minimal effort required to transform one probability distribution into another. It is widely used to evaluate how well synthetic (or anonymized) data preserves the statistical properties of real data (Qian et al., 2024; Sella et al., 2025). A lower Wasserstein distance indicates higher fidelity, which is often critical for downstream performance.

Figure 8 illustrates the results. As ϵ decreases below 1 (indicating stronger privacy), DP-kSame exhibits a sharp increase in Wasserstein distance, reflecting large deviations in the anonymized centroids from the real data. In contrast, DP-CVAE maintains a consistently low Wasserstein distance, even at $\epsilon = 0.01$, demonstrating robustness to tight privacy constraints.

This discrepancy stems from the mechanisms used. DP-kSame directly adds noise to centroids, which replace all data points in a cluster, thus amplifying distortion.

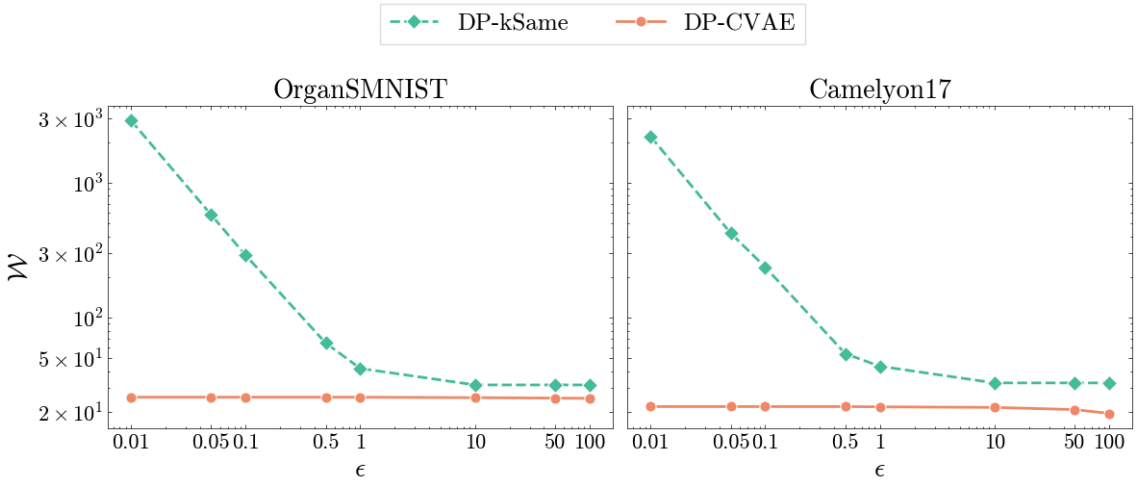

Figure 8: Privacy-utility trade-off: Wasserstein distance ($\mathcal{W}$), averaged across clients over three seed runs, between the distributions of real and shared embeddings versus DP privacy budget ϵ (lower ϵ = stronger privacy).

Conversely, DP-CVAE applies DP noise during model training and generates new samples without further perturbation, thereby preserving fidelity. The strong performance of DP-CVAE in the range of strict privacy budgets ($\epsilon \leq 1$) highlights its advantage in mitigating the privacy-utility trade-off in privacy-preserving FL.

4.3.2 Evaluation of Empirical Privacy

DP is the property of our data sharing mechanisms that establishes how protected individuals are when sharing their sensitive data. Some works (Wagner and Eckhoff, 2018) consider DP, along with k-anonymity, as a privacy metric. However, interpreting its parameters (*e.g.,* ϵ) can be difficult in practice. To offer a more intuitive and empirical perspective, multiple privacy metrics (PMs) have been proposed (see *e.g.,* Trudslev et al. (2025) and Wagner and Eckhoff (2018) for a comprehensive overview). We complement the theoretical DP guarantees with four PMs: two distance-based (so-called Unidentifiability Score and Distance to Closest Record (DCR)) and two attack-based (MIA and DOMIAS privacy attacks). Additionally, Unidentifiability Score and MIA F1-score measure record-level privacy, whereas DCR and DOMIAS F1-score offer a distributional view on privacy leakage. We follow existing implementations from Synthcity (Qian et al., 2023) and Syntheval (Lautrup et al., 2024) frameworks to compute these metrics.

Unidentifiability Score This score measures whether the anonymized records (including synthetic ones) are "different enough" from their real counterparts. Following Yoon et al. (2020), we compute *identifiability score* – the probability that an anonymized data point is closer to a real training point than that real point's nearest neighbor. The derived privacy score is then defined as 1 − identifiability score, where a score near 1 indicates strong privacy (minimal re-identification risk), and a score near 0 indicates poor protection.

Distance to Closest Record (DCR) DCR measures the median distance from each anonymized record to its nearest real record, normalized by distances among the real records (Steier et al., 2025; Trudslev et al., 2025). A high DCR suggests that anonymized samples are well-separated from real data, reducing memorization risk. To improve interpretability, we compare the DCR between anonymized and training data against a holdout set. If anonymized data is significantly closer to the training data than to the holdout data, we define a *DCR privacy loss*, meaning some information specific to the training data has leaked into the anonymized dataset. A negative DCR privacy loss suggests information leakage, while a positive one implies accuracy or fidelity loss.

MIA MIAs simulate an adversary attempting to determine whether a specific record was used in training. Following Syntheval's implementation (Lautrup et al., 2024), we adopt a black-box MIA setup using a Random Forest classifier to distinguish between training and holdout samples based on anonymized data (available to the attacker). Performance is evaluated via typical classification metrics, especially ROC-AUC and F1-score. Poor classification performance indicates strong privacy protection, as the attacker cannot confidently infer if a target record was a training member.

DOMIAS DOMIAS (van Breugel et al., 2023) is an unsupervised, density-based MIA that detects local overfitting via kernel density estimation. It estimates whether a data point is more likely to belong to the training distribution than a reference distribution. As with MIA, we compute metrics like ROC-AUC and F1-score; poorer performance indicates stronger privacy. While classifier-based MIAs might underestimate leakage if a poor classifier is chosen, DOMIAS is independent of model performance and captures distributional overexposure instead.

All PMs are computed directly on real embedding datasets as well as on anonymized datasets after they have been generated. Like prior work (Steier et al., 2025; van Breugel et al., 2023), we argue that such post-hoc, black-box metrics are more realistic for our FL setting, where clients only release anonymized or synthetic data, not raw inputs or models.

Figure 9 summarizes the trade-offs evaluated on Camelyon17 and OrganSMNIST datasets (not differentiating between IID and non-IID settings due to similar results). Both DP-kSame and DP-CVAE achieve strong privacy across all evaluation metrics. The unidentifiability scores for both methods are consistently close to 1.0, indicating that individual samples cannot be reliably traced back to their source. They both show less severe DCR privacy loss (higher values towards the positive side) than the Centroids-only baseline. Furthermore, the ROC-AUC values for both MIA and DOMIAS attacks (not included in the Figure) are close to 0.5 on both datasets. This is considered a strong privacy signal, suggesting that adversaries gain no advantage over random guessing. All other metrics for these attacks lie on the lower side, with F1-scores well below 0.5 and both precision and recall < 0.3. Compared to the baseline, these privacy attacks are clearly less successful on the anonymized data generated in DP-kSame and DP-CVAE paradigms.

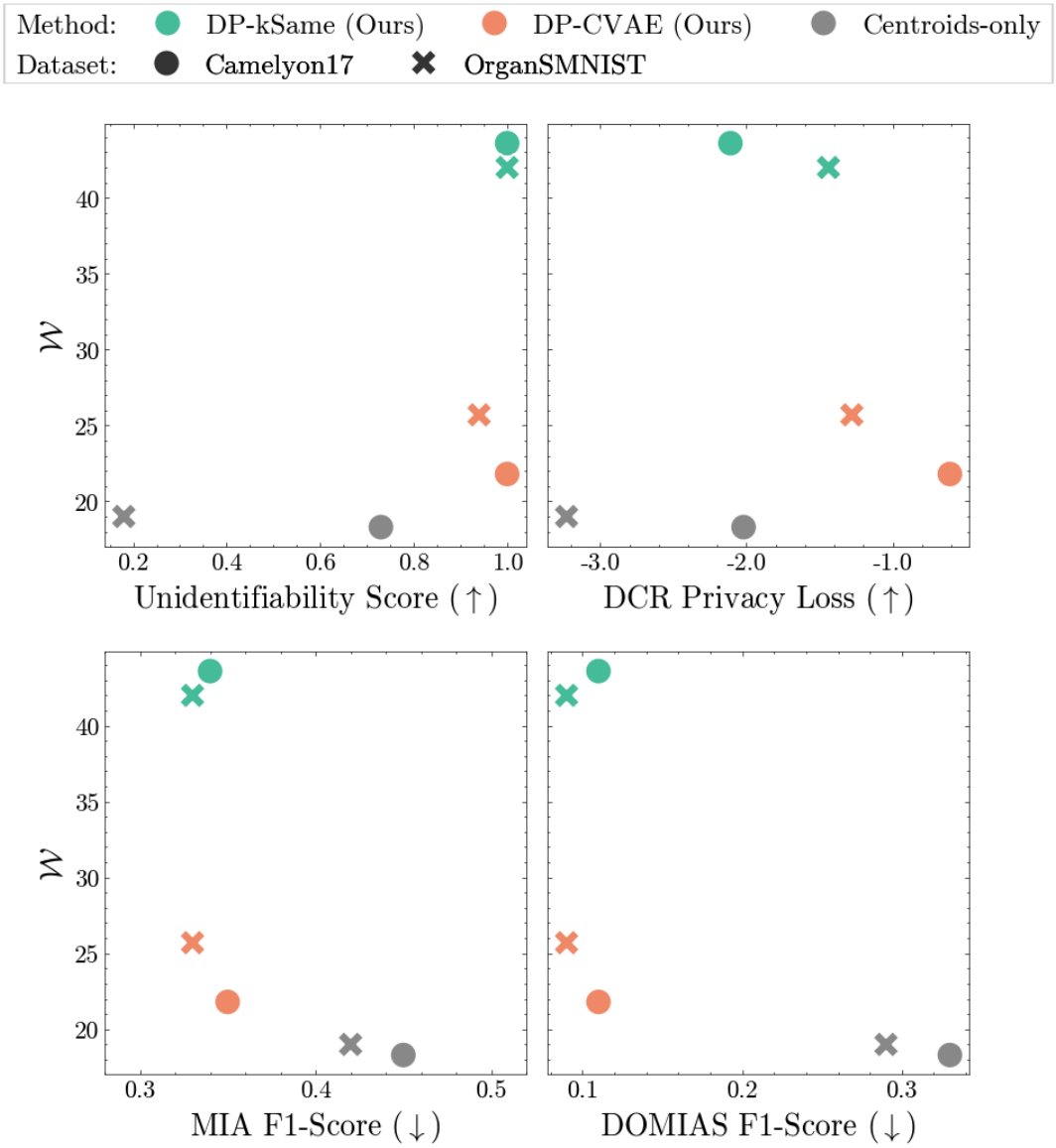

Figure 9: Privacy versus utility for DP-kSame and DP-CVAE, compared against a "Centroids-only" baseline (*i.e.,* no k-Same or DP). Privacy is measured Unidentifiability Score, DCR Privacy Loss, MIA F1-score, and DOMIAS F1-score; arrows indicate whether higher (↑) or lower (↓) values imply stronger privacy. Utility is measured by Wasserstein distance $\mathcal{W}$ (lower is better).

While DP-kSame and DP-CVAE offer comparable privacy, DP-kSame exhibits substantially higher Wasserstein distance, indicating reduced utility of its anonymized data. As discussed in 4.3.1, this degradation arises from the direct injection of noise into cluster centroids, which are then used to represent all samples in a group. The lack of a decoding mechanism to project these noisy centroids back into realistic sample space (based on a prior distribution) further limits fidelity. In contrast, DP-CVAE, though trained under similar DP constraints, can generate samples that remain statistically close to the real distribution. These findings highlight DP-

CVAE's strength in balancing privacy and utility. Its ability to maintain high generation fidelity while providing strong privacy guarantees makes it a more effective and scalable solution for privacy-preserving data sharing in FL.

4.4 Selection of Generative Models

We conduct an ablation study comparing DP-CVAE and DP-CGAN – two federated, differentially private generative models. Our goal is to evaluate their effectiveness in preserving the original data distribution, as well as their computational efficiency.

We measure fidelity using the average 2-Wasserstein distance ($\mathcal{W}$) between each client's real and synthetic datasets. As shown in Figure 10, DP-CVAE consistently achieves lower $\mathcal{W}$, indicating better approximation of the original data distribution.

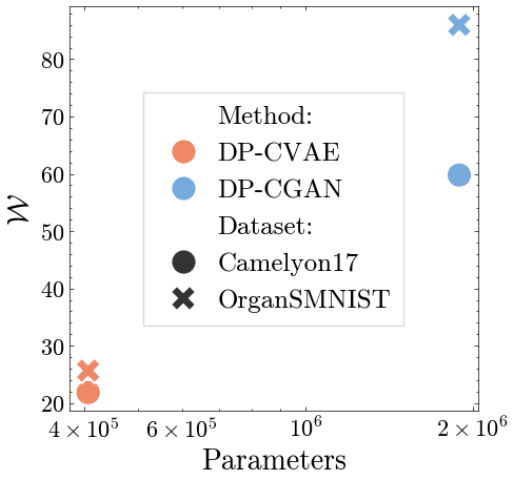

Figure 10: Fidelity of synthetic data generated by DP-CVAE versus DP-CGAN to the original data distribution, measured by the Wasserstein distance (lower is better). Results are analyzed in relation to the number of model parameters.

Interestingly, despite this clear difference in generation fidelity, the downstream classification accuracy and balanced accuracy of DP-CGAN and DP-CVAE are quite similar (see Table 3). This can likely be attributed to the semantic richness of FM embeddings, which makes downstream classifiers resilient to moderate distortions in input data. Nevertheless, DP-CVAE's higher fidelity demonstrates more faithful generation under DP, which is critical for generalizing to more complex or sensitive downstream tasks.

Moreover, DP-CVAE requires approximately 5× fewer parameters than DP-CGAN, reducing communication and computational overhead, which are critical in FL settings. While both models implement the same DP mechanism, DP-CVAE emerges as a favorable choice because of its efficiency and superior reconstruction quality.

5 Discussion

5.1 Key Findings and Contributions

A core novelty of our work is the paradigm shift it proposes: instead of training models on raw input data and sharing privacy-sensitive model updates, we explore representation-level data sharing in the latent space of a fixed, pre-trained FM. By leveraging rich and generalizable embeddings from FMs like DINOv2, we decouple local downstream modeling from global collaboration. This allows for privacy-enhanced and computationally lightweight participation, eliminating the need for costly end-to-end model synchronization. At the same time, clients can operate independently with simple classifiers and still benefit from shared knowledge through anonymized or synthetic embeddings.

To this end, we propose and evaluate two approaches for privacy-preserving representation sharing: (i) DP-kSame, a direct anonymization strategy based on k-anonymity with additive DP noise, and (ii) DP-CVAE, a generative framework that learns to generate privacy-aware synthetic embeddings through federated, differentially private CVAEs. Across diverse experiments on different medical image datasets, our findings reveal several consistent patterns that shed light on the practical effectiveness and privacy strengths of the proposed strategies.

In both of our proposed methods, downstream training remains private, personalized, and decoupled from the global collaboration. We demonstrate that, with this setup, both methods deliver stable performance across different FM backbones, indicating their general applicability for varied feature representations.

Among the evaluated methods, DP-CVAE emerges as the most effective privacy-preserving approach, offering the best balance between privacy protection and utility. In terms of utility, DP-CVAE consistently achieves competitive classification accuracy and reconstruction fidelity (measured by Wasserstein distance) compared to the baselines. On the privacy front, DP-CVAE performs best across all evaluated metrics. Its synthetic embeddings exhibit both lower record-level re-identifiability (measured by Unidentifiability Score) and reduced distributional memorization (measured by DCR Privacy Loss). They are also less vulnerable to both supervised (MIA) and unsupervised (DOMIAS) membership inference attacks. Notably, DP-CVAE outperforms DP-kSame in the privacy-utility trade-off under stringent privacy budgets (*e.g.,* $\epsilon \leq 1$). This advantage is attributable to the CVAE's ability to approximate the original data distribution while enforcing DP guarantees. Unlike direct anonymization through noise-perturbed centroids, DP-CVAE employs a generative modeling process that is more resilient to DP noise. Furthermore, by providing a globally trained class-conditional decoder, the DP-CVAE pipeline facilitates greater adaptability across applications without retraining the generative model. Compared to DP-CGAN, DP-CVAE generates higher-fidelity embeddings while incurring substantially lower computational and communication costs. This further underscores the practicality and scalability of the DP-CVAE framework in FL contexts.

As an exploratory approach, DP-kSame offers a conceptually simple and interpretable mechanism for embedding anonymization. Grounded in classical k-anonymity principles, this method replaces real data with cluster centroids and enforces formal privacy by adding DP noise before sharing these centroids. In classification tasks, DP-kSame performs on par with DP-CVAE. This confirms the viability of prototype sharing to support downstream inference, as suggested in prototype-based FL literature, especially when the embedding spaces are well-clustered. However, compared to DP-CVAE, DP-kSame can be more data-hungry, as evidenced by its notably lower accuracy on Camelyon17 (see Table 3) and overall lower balanced accuracy on OrganSMNIST as the number of local training samples decreases (see Figure 7). Besides, its practice of maintaining and transmitting a global datastore introduces additional storage and communication overhead.

DP-kSame also faces challenges in balancing privacy and utility, as its sensitivity to DP noise causes sharp fidelity degradation. Nevertheless, DP-kSame shows clear potential under looser privacy budgets and in empirical privacy evaluations, where its performance approaches that of DP-CVAE. Importantly, DP-kSame provides a transparent and computationally lightweight anonymization baseline that serves as an intermediary step between direct feature-sharing and generative synthesis. It helps us reason and motivate the need for a more flexible and noise-resilient approach like DP-CVAE.

In summary, our findings validate the effectiveness of privacy-preserving data sharing via pre-trained FM embeddings in FL. They demonstrate that DP-CVAE offers a scalable, robust and privacy-preserving solution, addressing multiple limitations of existing FL approaches. DP-kSame, while being less adaptive, indicates the viability of data anonymization through clustering and contributes to a better understanding of privacy–utility trade-offs in the embedding space. Together, these results chart a promising direction for designing data-centric FL solutions based on privacy-aware representation sharing.

5.2 Limitations and Future Work

While our work demonstrates the potential of privacy-preserving representation sharing in FL – particularly through our generative-based method, DP-CVAE – certain limitations remain, suggesting valuable directions for future research.

A key limitation arises in scenarios with highly heterogeneous data distributions and significant class imbalance. Although both DP-CVAE and DP-kSame generalize well in terms of accuracy, their balanced accuracy tend to suffer when minority classes are underrepresented in the synthetic or anonymized datasets. This stems from the globally aggregated nature of the shared data, which may not align with client-specific distributions. Consequently, global predictions can dilute local signals from underrepresented classes during interpolation.

To address this, future work could integrate ideas from long-tailed learning (Zhang et al., 2023), such as generative techniques designed to explicitly model minority

classes. Notably, our DP-CVAE currently samples embeddings with a fixed unit variance, allowing the model to learn adaptive or class-specific variance parameters could enhance the diversity and fidelity of generated data. Furthermore, conditioning the generative process on additional confounders, such as domain-specific attributes or client metadata, could enhance its ability to capture distributional nuances and mitigate inherent biases in the training distribution, thereby promoting generalization and fairness in real-world deployments.

Another interesting direction is to further investigate the potential of non-parametric models, such as k-NN, within the federated embedding sharing paradigm. In our analysis, k-NN classifiers generally underperform linear models, but in certain settings, especially under extreme non-IID conditions, they show competitive results. In general, their strengths lie in training-free deployment, adaptability (Imam et al., 2024; Marfoq et al., 2022), and resilience against catastrophic forgetting (Doerrich et al., 2024; Nakata et al., 2022). These properties suggest that k-NN could be revisited in scenarios where it may provide unique advantages, such as continual learning or edge deployment, where retraining is costly or infeasible.

The use of more specialized foundation models (FMs) also merits further investigation. We adopt DINOv2 for its strong general-purpose representations, but domain-specific FMs, such as those pre-trained on medical images, could offer improved performance in specialized tasks (Zhang et al., 2022; Zhou et al., 2023). Evaluating how such models interact with our privacy-preserving methods could enhance utility in high-stakes fields. Moreover, our current setup assumes all clients share a common FM backbone for feature extraction. In practice, however, clients may rely on heterogeneous FMs due to privacy concerns, resource constraints, or institutional differences. Future work could examine how this heterogeneity affects global representation alignment and downstream performance in our framework.

Finally, while our approach advances the trade-off between privacy, utility, and efficiency in FL through embedding-level sharing under formal privacy guarantees, further research is needed to evaluate broader concerns such as fairness, interpretability, and integration with other privacy-enhancing technologies. Continued efforts at the intersection of privacy, generative modeling, and federated systems design hold strong potential to extend and strengthen the framework proposed in this work.

6 Conclusion

This thesis proposes a new direction for privacy-preserving FL by shifting from traditional (downstream) model sharing to collaboration through differentially private representations in the rich embedding space of pre-trained FMs. We introduce and evaluate two novel methods within this paradigm: DP-kSame and DP-CVAE. DP-kSame serves as an interpretable intermediary between prototype sharing and synthetic data generation, helping to illuminate privacy–utility trade-offs in anonymized feature spaces. DP-CVAE, in contrast, leverages a generative approach that offers substantial improvements over DP-kSame and other baselines in balancing privacy and performance. Unlike conventional FL approaches that are tied to a specific downstream task, DP-CVAE enables clients to flexibly generate globally representative yet privacy-aware datasets for diverse applications.

Our experiments on medical imaging datasets show that both methods outperform end-to-end federated classifiers in accuracy while strengthening both theoretical and empirical privacy guarantees. Notably, DP-CVAE achieves higher data fidelity than DP-kSame and DP-CGAN, with lower communication and computation overhead. Overall, our proposed methods advance a growing trend in FL toward personalized, non-parametric, and representation-based collaboration (Collins et al., 2021; Marfoq et al., 2022; McLaughlin and Su, 2024; Tan et al., 2022; Tran et al., 2024), by addressing not only downstream performance but also privacy and communication – two central challenges in federated systems (Kairouz et al., 2021). Our findings position the proposed paradigm shift and especially DP-CVAE as a promising alternative for enabling secure, efficient, and adaptable FL workflows, particularly in privacy-sensitive domains such as medical imaging.

Bibliography

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016.

Idan Achituve, Aviv Shamsian, Aviv Navon, Gal Chechik, and Ethan Fetaya. Personalized federated learning with gaussian processes. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 8392–8406. Curran Associates, Inc., 2021.

Mohammed Adnan, Shivam Kalra, Jesse C. Cresswell, Graham W. Taylor, and Hamid R. Tizhoosh. Federated learning and differential privacy for medical image analysis. *Scientific Reports*, 12(1):1953, 2022. doi: 10.1038/s41598-022-05539-7.

Charu C. Aggarwal. On k-anonymity and the curse of dimensionality. In *Proceedings of the 31st International Conference on Very Large Data Bases*, VLDB '05, page 901–909. VLDB Endowment, 2005. ISBN 1595931546.

Fouzia Altaf, Syed M. S. Islam, Naveed Akhtar, and Naeem Khalid Janjua. Going deep in medical image analysis: Concepts, methods, challenges, and future directions. *IEEE Access*, 7:99540–99572, 2019. doi: 10.1109/ACCESS.2019.2929365.

George Annas. HIPAA regulations - A new era of medical-record privacy? *The New England journal of medicine*, 348:1486–90, 05 2003. doi: 10.1056/NEJMlim035027.

Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Singh, and Sunav Choudhary. Federated learning with personalization layers, 2019.

Zahra Azizi, Chao Zheng, Laura Mosquera, Louise Pilote, Khaled El Emam, and GOING-FWD Collaborators. Can synthetic data be a proxy for real clinical trial data? a validation study. *BMJ Open*, 11(4):e043497, Apr 2021. doi: 10.1136/bmjopen-2020-043497.

Aloïs Babé, Rémi Cuingnet, Mihaela Scuturici, and Serge Miguet. Generalization abilities of foundation models in waste classification. *Waste Management*, 198: 187–197, 05 2025. doi: 10.1016/j.wasman.2025.02.032.

Brett K. Beaulieu-Jones, Z. Steven Wu, Christopher Williams, Ran Lee, Sriram P. Bhavnani, James B. Byrd, and Casey S. Greene. Privacy-preserving generative deep neural networks support clinical data sharing. *Circulation: Cardiovascular Quality and Outcomes*, 12(7):e005122, Jul 2019. doi: 10.1161/CIRCOUTCOMES.118.005122.

Justin Brickell and Vitaly Shmatikov. The cost of privacy: Destruction of data-mining utility in anonymized data publishing. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, page 70–78, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605581934. doi: 10.1145/1401890.1401904.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33*, 2020.

Tianshi Cao, Alex Bie, Karsten Kreis, and Sanja Fidler. Differentially private generative models through optimal transport, 2021.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640, 2021. doi: 10.1109/ICCV48922.2021.00951.

James Casaletto, Alexander Bernier, Robyn McDougall, and Melissa S Cline. Federated analysis for privacy-preserving data sharing: A technical and legal primer. *Annual Review of Genomics and Human Genetics*, 24:347–368, 2023. doi: 10.1146/annurev-genom-110122-084756.

Baojian Chen, Hongjia Li, Lu Guo, and Liming Wang. Label-wise distribution adaptive federated learning on non-iid data. In *2023 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1–6, 2023a.

Bingyang Chen, Tao Chen, Xingjie Zeng, Weishan Zhang, Qinghua Lu, Zhaoxiang Hou, Jiehan Zhou, and Sumi Helal. DFML: Dynamic federated meta-learning for rare disease prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 21(4):880–889, 2024. doi: 10.1109/TCBB.2023.3239848.

Hong-You Chen and Wei-Lun Chao. On bridging generic and personalized federated learning for image classification. In *International Conference on Learning Representations*, 2022.

Minghui Chen, Meirui Jiang, Qi Dou, Zehua Wang, and Xiaoxiao Li. FedSoup: improving generalization and personalization in federated learning via selective model interpolation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 318–328. Springer, 2023b.

Yao Chen, Yijie Gui, Hong Lin, Wensheng Gan, and Yongdong Wu. Federated learning attacks and defenses: A survey. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 4256–4265, Los Alamitos, CA, USA, December 2022. IEEE Computer Society. doi: 10.1109/BigData55660.2022.10020431.

Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F. Stewart, and Jimeng Sun. Generating multi-label discrete patient records using generative adversarial networks. In Finale Doshi-Velez, Jim Fackler, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 of *Proceedings of Machine Learning Research*, pages 286–305. PMLR, 18–19 Aug 2017.

Aloni Cohen and Kobbi Nissim. Towards formalizing the GDPR's notion of singling out. *Proceedings of the National Academy of Sciences*, 117(15):8344–8352, 2025/07/06 2020. doi: 10.1073/pnas.1914598117.

Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2089–2099. PMLR, 18–24 Jul 2021.

Teodora Curelariu and Alexandre Lodie. The lawfulness of re-identification under data protection law. In *Privacy Technologies and Policy: 12th Annual Privacy Forum, APF 2024, Karlstad, Sweden, September 4–5, 2024, Proceedings*, page 112–131, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-68023-6. doi: 10.1007/978-3-031-68024-3_6.

Hari Prasanna Das, Ryan Tran, Japjot Singh, Xiangyu Yue, Geoffrey H. Tison, Alberto Sangiovanni-Vincentelli, and Costas Spanos. Conditional synthetic data generation for robust machine learning applications with limited pandemic data. In *epiDAMIK 5.0: The 5th International workshop on Epidemiology meets Data Mining and Knowledge discovery at KDD 2022*, 2022.

Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Adaptive personalized federated learning, 2020.

Francesco Di Salvo, David Tafler, Sebastian Doerrich, and Christian Ledig. Privacy-preserving datasets by capturing feature distributions with Conditional VAEs. *The 35th British Machine Vision Conference*, 2024.

Sebastian Doerrich, Tobias Archut, Francesco Di Salvo, and Christian Ledig. Integrating kNN with foundation models for adaptable and privacy-aware image classification. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, 2024. doi: 10.1109/ISBI56570.2024.10635560.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg

Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. 2024.

Yichao Du, Zhirui Zhang, Bingzhe Wu, Lemao Liu, Tong Xu, and Enhong Chen. Federated nearest neighbor machine translation. In *The Eleventh International Conference on Learning Representations*, 2023.

Cynthia Dwork. Differential privacy. In *Proceedings of the 33rd International Colloquium on Automata, Languages and Programming*, pages 1–12. Springer Berlin Heidelberg, 2006.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 01 2013. doi: 10.1561/0400000042.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Conference on Theory of Cryptography*, TCC'06, page 265–284, Berlin, Heidelberg, 2006. Springer-Verlag. ISBN 3540327312. doi: 10.1007/11681878_14.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3557–3568. Curran Associates, Inc., 2020.

Joao Fonseca and Fernando Bação. Tabular and latent space synthetic data generation: a literature review. *Journal of Big Data*, 10, 07 2023. doi: 10.1186/s40537-023-00792-7.

Luca Frigerio, Alexandre S. de Oliveira, Lluis Gomez, and Pascal Duverger. Differentially private generative adversarial networks for time series, continuous, and discrete open data. In Gurpreet Dhillon, Fredrik Karlsson, Karin Hedström, and André Zúquete, editors, *ICT Systems Security and Privacy Protection. SEC 2019*, volume 562 of *IFIP Advances in Information and Communication Technology*, pages 151–164. Springer, Cham, 2019. doi: 10.1007/978-3-030-22312-0_11.

Liang Gao, Huazhu Fu, Li Li, Yingwen Chen, Ming Xu, and Cheng-Zhong Xu. FedDC: Federated learning with non-IID data via local drift decoupling and correction. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10102–10111, 2022. doi: 10.1109/CVPR52688.2022.00987.

Sonakshi Garg and Vicenç Torra. Privacy in manifolds: Combining k-anonymity with differential privacy on fréchet means. *Computers & Security*, 144:103983, 2024. doi: https://doi.org/10.1016/j.cose.2024.103983.

Ashkan Vedadi Gargary and Emiliano De Cristofaro. A systematic review of federated generative models. 2024.

Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1732–1742, 2020.

Mauro Giuffrè and Dennis L. Shung. Harnessing the power of synthetic data in healthcare: Innovation, application, and privacy. *npj Digital Medicine*, 6(1):186, 2023. doi: 10.1038/s41746-023-00927-3.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Ehsan Hallaji, Roozbeh Razavi-Far, Mehrdad Saif, Boyu Wang, and Qiang Yang. Decentralized federated learning: A survey on security and privacy. *IEEE Transactions on Big Data*, 10:194–213, 04 2024. doi: 10.1109/TBDATA.2024.3362191.

Mozafar Hayaeian Shirvan, Mohammad Hossein Moattar, and Mehdi Hosseinzadeh. Deep generative approaches for oversampling in imbalanced data classification problems: A comprehensive review and comparative analysis. *Applied Soft Computing*, 170:112677, 2025. doi: https://doi.org/10.1016/j.asoc.2024.112677.

Peilin He, Chenkai Lin, and Isabella Montoya. DPFedBank: Crafting a privacy-preserving federated learning framework for financial institutions with policy pillars, 10 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Naoise Holohan, Spiros Antonatos, Stefano Braghin, and Pól Mac Aonghusa. (k,ϵ)-anonymity: k-anonymity with ϵ-differential privacy, 2017.

Fariha Imam, Petr Musilek, and Marek Z. Reformat. Parametric and nonparametric machine learning techniques for increasing power system reliability: A review. *Information*, 15(1), 2024. doi: 10.3390/info15010037.

Malhar Jere, Tyler Farnan, and Farinaz Koushanfar. A taxonomy of attacks on federated learning. *IEEE Security & Privacy*, 19:20–28, 2021. doi: 10.1109/MSEC.2020.3039941.

Yifan Jiang, Haofu Chen, Murray Loew, and Hanbin Ko. COVID-19 CT image synthesis with a conditional generative adversarial network. *IEEE Journal of Biomedical and Health Informatics*, 25(2):441–452, Feb 2021. doi: 10.1109/JBHI.2020.3042523.

James Jordon, Jinsung Yoon, and Mihaela van der Schaar. PATE-GAN: Generating synthetic data with differential privacy guarantees. In *International Conference on Learning Representations*, 2018.

Bayrem Kaabachi, Jérémie Despraz, Thierry Meurers, Karen Otte, Mehmed Halilovic, Bogdan Kulynych, Fabian Prasser, and Jean Louis Raisaro. A scoping review of privacy and utility metrics in medical synthetic data. *npj Digital Medicine*, 8(1):60, 2025. doi: 10.1038/s41746-024-01359-3.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawit, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konecný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Hang Qi, Daniel Ramage, Ramesh Raskar, Mariana Raykova, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning, 2021.

Georgios Kaissis, Alexander Ziller, Jonathan Passerat-Palmbach, Théo Ryffel, Dmitrii Usynin, Andrew Trask, Ionésio Lima, Jason Mancuso, Friederike Jungmann, Marc-Matthias Steinborn, Andreas Saleh, Marcus Makowski, Daniel Rueckert, and Rickmer Braren. End-to-end privacy preserving deep learning on multi-institutional medical imaging. *Nature Machine Intelligence*, 3(6):473–484, 2021. doi: 10.1038/s42256-021-00337-8.

Sai Praneeth Karimireddy, Wenshuo Guo, and Michael I. Jordan. Mechanisms that incentivize data sharing in federated learning, 2022.

Shamisa Kaspour and Abdulsalam Yassine. Variational auto-encoder model and federated approach for non-intrusive load monitoring in smart homes. In *2023 IEEE Symposium on Computers and Communications (ISCC)*, pages 1110–1115, 2023.

Rémi Kazmierczak, Eloïse Berthier, Goran Frehse, and Gianni Franchi. Explainability and vision foundation models: A survey. *Information Fusion*, 122:103184, 2025. doi: https://doi.org/10.1016/j.inffus.2025.103184.

Afsana Khan, Marijn ten Thij, and Anna Wilbik. Vertical Federated Learning: A Structured Literature Review. *Knowledge and Information Systems*, 67(4):3205–3243, 2025. doi: 10.1007/s10115-025-02356-y.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014*, 2014.

Lennart R Koetzier, Jie Wu, Domenico Mastrodicasa, Aline Lutz, Matthew Chung, W Adam Koszek, Jayanth Pratap, Akshay S Chaudhari, Pranav Rajpurkar, Matthew P Lungren, et al. Generating synthetic data for medical imaging. *Radiology*, 312(3):e232471, 2024.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, 2021.

Vaijainthymala Krishnamoorthy. Data obfuscation through latent space projection for privacy-preserving AI governance: Case studies in medical diagnosis and finance fraud detection. *JMIRx Med*, 6:e70100, Mar 2025. doi: 10.2196/70100.

Ira Ktena, Olivia Wiles, Isabela Albuquerque, Sylvestre-Alvise Rebuffi, Ryutaro Tanno, Abhijit Guha Roy, Shekoofeh Azizi, Danielle Belgrave, Pushmeet Kohli, Taylan Cemgil, et al. Generative models improve fairness of medical classifiers under distribution shifts. *Nature Medicine*, 30(4), 2024.

Lucas Lange, Maja Schneider, Peter Christen, and Erhard Rahm. Privacy in practice: Private COVID-19 detection in X-Ray images (extended version). *arXiv:2211.11434*, 2022.

Stefan Larson, Anthony Zheng, Anish Mahendran, Rishi Tekriwal, Adrian Cheung, Eric Guldan, Kevin Leach, and Jonathan K. Kummerfeld. Iterative feature mining for constraint-based data collection to increase data diversity and model robustness. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8097–8106, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.650.

Anton D. Lautrup, Tobias Hyrup, Arthur Zimek, and Peter Schneider-Kamp. Syntheval: a framework for detailed utility and privacy evaluation of tabular synthetic data. *Data Mining and Knowledge Discovery*, 39(1), 2024. doi: 10.1007/s10618-024-01081-4.

Daliang Li and Junpu Wang. FedMD: Heterogenous federated learning via model distillation, 2019.

Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-Closeness: Privacy beyond k-Anonymity and l-Diversity. In *2007 IEEE 23rd International Conference on Data Engineering*, pages 106–115, 2007. doi: 10.1109/ICDE.2007.367856.

Shenghui Li, Fanghua Ye, Meng Fang, Jiaxu Zhao, Yun-Hin Chan, Edith C. H. Ngai, and Thiemo Voigt. Synergizing foundation models and federated learning: A survey, 2024a.

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.

Xuyang Li, Weizhuo Zhang, Yue Yu, Wei-Shi Zheng, Tong Zhang, and Ruixuan Wang. SiFT: A serial framework with textual guidance for federated learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2024b.

Bill Yuchen Lin, Chaoyang He, Zihang Ze, Hulin Wang, Yufen Hua, Christophe Dupuy, Rahul Gupta, Mahdi Soltanolkotabi, Xiang Ren, and Salman Avestimehr. FedNLP: Benchmarking federated learning methods for natural language processing tasks. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 157–175, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.13.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

Haiyang Liu, Yuki Endo, Jinho Lee, and Shunsuke Kamijo. PREmbed: Balancing conditional generative models with embedding pretraining and regularization. *Electronics*, 14(2), 2025. doi: 10.3390/electronics14020280.

Liyuan Liu and Meng Han. Data sharing and exchanging with incentive and optimization: A survey. *Discover Data*, 2(1):2, 2024. doi: 10.1007/s44248-024-00006-2.

Tao Liu, Zhi Wang, Hui He, Wei Shi, Liangliang Lin, Ran An, and Chenhao Li. Efficient and secure federated learning for financial applications. *Applied Sciences*, 13(10), 2023. doi: 10.3390/app13105877.

Xiaonan Liu, Yansha Deng, Arumugam Nallanathan, and Mehdi Bennis. Federated learning and meta learning: Approaches, applications, and directions. *IEEE Communications Surveys Tutorials*, 26(1):571–618, 2024. doi: 10.1109/COMST.2023.3330910.

Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, 1(1):3–es, March 2007. ISSN 1556-4681. doi: 10.1145/1217299.1217302.

Yishay Mansour, Mehryar Mohri, Jae Ro, and Ananda Theertha Suresh. Three approaches for personalization with applications to federated learning, 2020.

Yuzhu Mao, Zihao Zhao, Guangfeng Yan, Yang Liu, Tian Lan, Linqi Song, and Wenbo Ding. Communication-efficient federated learning with adaptive quantization. 13(4), August 2022. ISSN 2157-6904. doi: 10.1145/3510587.

Othmane Marfoq, Giovanni Neglia, Richard Vidal, and Laetitia Kameni. Personalized federated learning through local memorization. In *International Conference on Machine Learning*, pages 15070–15092. PMLR, 2022.

Connor J. McLaughlin and Lili Su. Personalized federated learning via feature distribution adaptation. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 77038–77059. Curran Associates, Inc., 2024.

Brendan McMahan and Daniel Ramage. Federated learning: Collaborative machine learning without centralized training data, 2017. URL `https://ai.googleblog.com/2017/04/federated-learning-collaborative.html`. Google AI Blog.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, 2017.

Blaž Meden, Žiga Emeršič, Vitomir Štruc, and Peter Peer. k-Same-Net: k-Anonymity with generative deep neural networks for face deidentification. *Entropy*, 20(1), 2018. ISSN 1099-4300. doi: 10.3390/e20010060.

Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 691–706, 2019. doi: 10.1109/SP.2019.00029.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets, 2014.

Mahdi Morafah, Weijia Wang, and Bill Lin. A practical recipe for federated learning under statistical heterogeneity experimental design. *IEEE Transactions on Artificial Intelligence*, 5(4):1708–1717, 2024. doi: 10.1109/TAI.2023.3297090.

Praveen Kumar Myakala, Chiranjeevi Bura, and Anil Kumar Jonnalagadda. Federated learning and data privacy: A review of challenges and opportunities. *International Journal of Research Publication and Reviews*, 5:1867–1879, 12 2024. doi: 10.55248/gengpi.5.1224.3512.

Kengo Nakata, Youyang Ng, Daisuke Miyashita, Asuka Maki, Yu-Chieh Lin, and Jun Deguchi. Revisiting a kNN-based image classification system with high-capacity storage. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision –ECCV 2022*, pages 457–474, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19836-6.

Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125, 2008. doi: 10.1109/SP.2008.33.

Mahmood Nasr, Shuang Song, Abhradeep Thakurta, Nicolas Papernot, and Nicholas Carlini. Adversary instantiation: Lower bounds for differentially private machine learning. In *2021 IEEE Symposium on Security and Privacy (S&P)*. IEEE, 2021.

E.M. Newton, L. Sweeney, and B. Malin. Preserving privacy by de-identifying face images. *IEEE Transactions on Knowledge and Data Engineering*, 17(2):232–243, 2005. doi: 10.1109/TKDE.2005.32.

Kobbi Nissim. Privacy: From database reconstruction to legal theorems. In *Proceedings of the 40th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, PODS'21, page 33–41, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383813. doi: 10.1145/3452021.3458816.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Ranadeep Reddy Palle. Synthetic data generation for privacy-preserving machine learning training. *International Journal of Research and Analytical Reviews*, 5, 01 2018.

Victor M. Panaretos and Yoav Zemel. Statistical aspects of wasserstein distances. *Annual Review of Statistics and Its Application*, 6(1):405–431, March 2019. ISSN 2326-831X. doi: 10.1146/annurev-statistics-030718-104938.

Connie L. Parks and Keith L. Monson. Automated facial recognition of computed tomography-derived facial images: Patient privacy implications. *Journal of Digital Imaging*, 30(2):204–214, 2017. doi: 10.1007/s10278-016-9932-7.

Sayak Paul and Pin-Yu Chen. Vision transformers are robust learners. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(2):2071–2081, Jun. 2022. doi: 10.1609/aaai.v36i2.20103.

Bjarne Pfitzner and Bert Arnrich. Dpd-fvae: Synthetic data generation using federated variational autoencoders with differentially-private decoder. *arXiv:2211.11591*, 2022.

Zhaozhi Qian, Bogdan-Constantin Cebere, and Mihaela van der Schaar. Synthcity: facilitating innovative use cases of synthetic data in different data modalities, 2023.

Zhaozhi Qian, Thomas Callender, Bogdan Cebere, Sam M. Janes, Neal Navani, and Mihaela van der Schaar. Synthetic data for privacy-preserving clinical risk prediction. *Scientific Reports*, 14(1):25676, 2024. doi: 10.1038/s41598-024-72894-y.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, page 8748–8763. PMLR, 2021.

Ashish Rauniyar, Desta Haileselassie Hagos, Debesh Jha, Jan Erik Håkegård, Ulas Bagci, Danda B. Rawat, and Vladimir Vlassov. Federated learning for medical applications: A taxonomy, current trends, challenges, and future research directions. *IEEE Internet of Things Journal*, 11(5):7374–7398, 2024. doi: 10.1109/JIOT.2023.3329061.

Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletarì, Holger R. Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N. Galtier, Bennett A. Landman, Klaus Maier-Hein, Sébastien Ourselin, Micah Sheller, Ronald M. Summers, Andrew Trask, Daguang Xu, Maximilian Baust, and M. Jorge Cardoso. The future of digital health with federated learning. *npj Digital Medicine*, 3(1):119, 2020. doi: 10.1038/s41746-020-00323-1.

Nuria Rodríguez-Barroso, Daniel Jiménez-López, M. Victoria Luzón, Francisco Herrera, and Eugenio Martínez-Cámara. Survey on federated learning threats: Concepts, taxonomy on attacks and defences, experimental study and challenges. *Information Fusion*, 90:148–173, February 2023. ISSN 1566-2535. doi: 10.1016/j.inffus.2022.09.011.

Nawraz Saeed, Mohamed Ashour, and Maggie Mashaly. Comprehensive review of federated learning challenges: a data preparation viewpoint. *Journal of Big Data*, 12, 06 2025. doi: 10.1186/s40537-025-01195-6.

Felix Sattler, Klaus-Robert Müller, and Wojciech Samek. Clustered Federated Learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE Transactions on Neural Networks and Learning Systems*, 32(8): 3710–3722, 2020. doi: 10.1109/TNNLS.2020.3015958.

Christopher Schwarz, Walter Kremers, Terry Therneau, Richard Sharp, Jeffrey Gunter, Prashanthi Vemuri, Arvin Arani, Anthony Spychalla, Kejal Kantarci, David Knopman, Ronald Petersen, and Clifford Jack. Identification of anonymous MRI research participants with face-recognition software. *New England Journal of Medicine*, 381:1684–1686, 10 2019. doi: 10.1056/NEJMc1908881.

Jonathan Scott, Hossein Zakerinia, and Christoph H Lampert. PeFLL: Personalized federated learning by learning to learn. In *The Twelfth International Conference on Learning Representations*, 2024.

Nadir Sella, Florent Guinot, Nikita Lagrange, Laurent-Philippe Albou, Jonathan Desponds, and Hervé Isambert. Preserving information while respecting privacy through an information theoretic framework for synthetic health data generation. *npj Digital Medicine*, 8(1):49, 2025. doi: 10.1038/s41746-025-01431-6.

Yueyue Shi, Hengjie Song, and Jun Xu. Responsible and effective federated learning in financial services: A comprehensive survey. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pages 4229–4236, 2023. doi: 10.1109/CDC49753.2023.10384119.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in neural information processing systems*, pages 3483–3491, 2015.

Amy Steier, Lipika Ramaswamy, Andre Manoel, and Alexa Haushalter. Synthetic data privacy metrics, 2025.

Fahim Sufi. Addressing data scarcity in the medical domain: A gpt-based approach for synthetic data generation and feature extraction. *Information*, 15(5), 2024. doi: 10.3390/info15050264.

Chang Sun, Johan Soest, and Michel Dumontier. Generating synthetic personal health data using conditional generative adversarial networks combining with differential privacy. *Journal of Biomedical Informatics*, 143:104404, 06 2023. doi: 10.1016/j.jbi.2023.104404.

Latanya Sweeney. k-Anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570, 2002. doi: 10.1142/S0218488502001648.

Alysa Ziying Tan, Han Yu, Lizhen Cui, and Qiang Yang. Towards personalized federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 34(12):9587–9603, December 2023. ISSN 2162-2388. doi: 10.1109/tnnls.2022.3160699.

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.

Reihaneh Torkzadehmahani, Peter Kairouz, and Benedict Paten. DP-CGAN: Differentially private synthetic data and label generation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 98–104, 2019. doi: 10.1109/CVPRW.2019.00018.

Trung-Kien Tran, Ha-Phuong Tran, Thi-Lan Le, and Thanh-Hai Tran. FedNTProto: A prototype-based approach for personalized federated learning. In *2024 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, pages 1–6, 2024. doi: 10.1109/MAPR63514.2024.10660707.

Frederik Marinus Trudslev, Matteo Lissandrini, Juan Manuel Rodriguez, Martin Bøgsted, and Daniele Dell'Aglio. A review of privacy metrics for privacy-preserving synthetic data generation, 2025.

Ioannis N Tzortzis, Alberto Gutierrez-Torre, Stavros Sykiotis, Ferran Agulló, Nikolaos Bakalos, Anastasios Doulamis, Nikolaos Doulamis, and Josep Ll Berral. Towards generalizable federated learning in medical imaging: A real-world case study on mammography data. *Computational and Structural Biotechnology Journal*, 28: 106–107, 2025. doi: 10.1016/j.csbj.2025.03.031.

Boris van Breugel, Hao Sun, Zhaozhi Qian, and Mihaela van der Schaar. Membership inference attacks against synthetic data through overfitting detection. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 3493–3514. PMLR, 25–27 Apr 2023.

Paul Voigt and Axel von dem Bussche. *The EU General Data Protection Regulation (GDPR): A Practical Guide*. Springer Publishing Company, Incorporated, 1st edition, 2017. ISBN 3319579584.

Lahari Voleti and Shen Shyang Ho. Personalized learning with limited data on edge devices using federated learning and meta-learning. In *Proceedings of the Eighth ACM/IEEE Symposium on Edge Computing*, SEC '23, page 378–382, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400701238. doi: 10.1145/3583740.3626811. URL `https://doi.org/10.1145/3583740.3626811`.

Isabel Wagner and David Eckhoff. Technical privacy metrics: A systematic survey. *ACM Computing Survey*, 51(3), June 2018. ISSN 0360-0300. doi: 10.1145/3168389.

Pengju Wang, Bochao Liu, Weijia Guo, Yong Li, and Shiming Ge. Towards personalized federated learning via comprehensive knowledge distillation, 2024.

Zhenchen Wang, Puja Myles, and Allan Tucker. Generating and evaluating synthetic UK primary care data: Preserving data utility patient privacy. In *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 126–131, 2019. doi: 10.1109/CBMS.2019.00036.

Chuhan Wu, Fangzhao Wu, Lingjuan Lyu, Yongfeng Huang, and Xing Xie. Communication-efficient federated learning via knowledge distillation. *Nature Communications*, 13(1):2032, 2022. doi: 10.1038/s41467-022-29763-x.

Jin Xie, Chenqing Zhu, and Songze Li. Fedmes: Personalized federated continual learning leveraging local memory. *CoRR*, abs/2404.12710, 2024. URL `https://doi.org/10.48550/arXiv.2404.12710`.

Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. Differentially private generative adversarial network, 2018.

Xuanang Xu, Fugen Zhou, Bo Liu, Dongshan Fu, and Xiangzhi Bai. Efficient multiple organ localization in CT image using 3D region proposal network. *IEEE transactions on medical imaging*, 38(8):1885–1898, 2019.

Yikai Xu and Hongbo Fan. FedDK: Improving cyclic knowledge distillation for personalized healthcare federated learning. *IEEE Access*, 11:72409–72417, 2023. doi: 10.1109/ACCESS.2023.3294812.

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):41, 01 2023. doi: 10.1038/s41597-022-01721-8.

Mang Ye, Xiuwen Fang, Bo Du, Pong C. Yuen, and Dacheng Tao. Heterogeneous federated learning: State-of-the-art and research challenges. *ACM Computing Surveys*, 56(3), October 2023. ISSN 0360-0300. doi: 10.1145/3625558.

Jinsung Yoon, Lydia N. Drumright, and Mihaela van der Schaar. Anonymization through data synthesis using generative adversarial networks (ADS-GAN). *IEEE Journal of Biomedical and Health Informatics*, 24(8):2378–2388, 03 2020. doi: 10.1109/JBHI.2020.2980262.

Naoya Yoshida, Takayuki Nishio, Masahiro Morikura, Koji Yamamoto, and Ryo Yonetani. Hybrid-FL for wireless networks: Cooperative learning mechanism using non-IID data. *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*, pages 1–7, 2019.

Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, Graham Cormode, and Ilya Mironov. Opacus: User-friendly differential privacy library in PyTorch. 2022.

Yu Yu, Shahram Khadivi, and Jia Xu. Can data diversity enhance learning generalization? In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4933–4945, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

Betul Yurdem, Murat Kuzlu, Mehmet Kemal Gullu, Ferhat Ozgur Catak, and Maliha Tabassum. Federated Learning: Overview, Strategies, Applications, Tools and Future Directions. *Heliyon*, 10(19):e38137, 2024. doi: https://doi.org/10.1016/j.heliyon.2024.e38137.

Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, Nghia Hoang, and Yasaman Khazaeni. Bayesian nonparametric federated learning of neural networks, 09–15 Jun 2019.

Tuo Zhang, Tiantian Feng, Samiul Alam, Dimitrios Dimitriadis, Mi Zhang, Shrikanth Narayanan, and Salman Avestimehr. GPT-FL: Generative pre-trained model-assisted federated learning, 2024.

Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Yitian Zhang, Xu Ma, Yue Bai, Huan Wang, and Yun Fu. Accessing vision foundation models via ImageNet-1K. In *The Thirteenth International Conference on Learning Representations*, 2025.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. Contrastive learning of medical visual representations from paired images and text. In Zachary Lipton, Rajesh Ranganath, Mark Sendak, Michael Sjoding, and Serena Yeung, editors, *Proceedings of the 7th Machine Learning for Healthcare Conference*, volume 182 of *Proceedings of Machine Learning Research*, pages 2–25. PMLR, 05–06 Aug 2022.

Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-IID data. 2018. doi: 10.48550/ARXIV.1806.00582.

Yukun Zhou, Mark A. Chia, Siegfried K. Wagner, Murat S. Ayhan, Dominic J. Williamson, Robbert R. Struyven, Timing Liu, Moucheng Xu, Mateo G. Lozano, Peter Woodward-Court, Yuka Kihara, Naomi Allen, John E. J. Gallacher, Thomas Littlejohns, Tariq Aslam, Paul Bishop, Graeme Black, Panagiotis Sergouniotis, Denize Atan, Andrew D. Dick, Cathy Williams, Sarah Barman, Jenny H. Barrett, Sarah Mackie, Tasanee Braithwaite, Roxana O. Carare, Sarah Ennis, Jane Gibson, Andrew J. Lotery, Jay Self, Usha Chakravarthy, Ruth E. Hogg, Euan Paterson, Jayne Woodside, Tunde Peto, Gareth Mckay, Bernadette Mcguinness, Paul J. Foster, Konstantinos Balaskas, Anthony P. Khawaja, Nikolas Pontikos, Jugnoo S. Rahi, Gerassimos Lascaratos, Praveen J. Patel, Michelle Chan, Sharon Y. L. Chua, Alexander Day, Parul Desai, Cathy Egan, Marcus Fruttiger, David F. Garway-Heath, Alison Hardcastle, Sir Peng T. Khaw, Tony Moore, Sobha Sivaprasad, Nicholas Strouthidis, Dhanes Thomas, Adnan Tufail, Ananth C. Viswanathan, Bal Dhillon, Tom Macgillivray, Cathie Sudlow, Veronique Vitart, Alexander Doney, Emanuele Trucco, Jeremy A. Guggeinheim, James E. Morgan, Chris J. Hammond, Katie Williams, Pirro Hysi, Simon P. Harding, Yalin Zheng, Robert Luben, Phil Luthert, Zihan Sun, Martin McKibbin, Eoin O'Sullivan, Richard Oram, Mike Weedon, Chris G. Owen, Alicja R. Rudnicka, Naveed Sattar, David Steel, Irene Stratton, Robyn Tapp, Max M. Yates, Axel Petzold, Savita Madhusudhan, Andre Altmann, Aaron Y. Lee, Eric J. Topol, Alastair K. Denniston, Daniel C. Alexander, Pearse A. Keane, and UK Biobank Eye & Vision Consortium. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981):156–163, 2023. doi: 10.1038/s41586-023-06555-x.

Hangyu Zhu, Jinjin Xu, Shiqing Liu, and Yaochu Jin. Federated learning on non-IID data: A survey. *Neurocomputing*, 465, 09 2021. doi: 10.1016/j.neucom.2021.07.098.

Alexander Ziller, Dmitrii Usynin, Rickmer Braren, Marcus Makowski, Daniel Rueckert, and Georgios Kaissis. Medical imaging deep learning with differential privacy. *Scientific Reports*, 11(1):13524, 2021. doi: 10.1038/s41598-021-93030-0.

Alexander Ziller, Tamara Mueller, Simon Stieger, Leonhard Feiner, Johannes Brandt, Rickmer Braren, Daniel Rueckert, and Georgios Kaissis. Reconciling privacy and accuracy in ai for medical imaging. *Nature Machine Intelligence*, 6: 1–11, 06 2024. doi: 10.1038/s42256-024-00858-y.

Declaration of Authorship

Ich erkläre hiermit gemäß §9 Abs. 12 APO, dass ich die vorstehende Abschlussarbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Des Weiteren erkläre ich, dass die digitale Fassung der gedruckten Ausfertigung der Abschlussarbeit ausnahmslos in Inhalt und Wortlaut entspricht und zur Kenntnis genommen wurde, dass diese digitale Fassung einer durch Software unterstützten, anonymisierten Prüfung auf Plagiate unterzogen werden kann.

Place, Date	Signature