# Federated Chest X-ray Diagnosis under Label Heterogeneity via Disease-Aligned Prototype

Obed Jamir[1(✉)] and Angshuman Paul[1]

Indian Institute of Technology Jodhpur, Rajasthan, India
✉Correspondence: d23cse004@iitj.ac.in

**Abstract.** Chest x-ray diagnosis is performed across hospitals with diverse patient populations and clinical priorities, so different institutions often annotate different subsets of thoracic abnormalities. This variation in label distributions is inherent to medical federated learning (FL) and constrains its deployment in real-world clinical settings, yet remains largely underexplored. Most existing FL approaches assume a homogeneous label space or require institutions to share their diagnostic labels to construct a global taxonomy. These assumptions rarely hold in practice due to privacy regulations and institutional constraints. We propose FedDAP, a federated learning framework for chest x-ray diagnosis with heterogeneous labels that operates without sharing local labels or raw data. Our approach maps local feature representations into a shared semantic space defined by disease-aligned class prototypes. At the server, a dynamic layer-wise aggregation strategy weights each layer according to its discriminative capability, allowing clients to contribute meaningfully across the network layers. Experiments on three public chest x-ray datasets under 3-client and 6-client configurations show consistent gains over the strongest state-of-the-art baseline, with a relative improvement of up to 3.7% in accuracy and up to +1.60 AUROC. The code is available at https://github.com/obedjamir/FedDAP.git.

**Keywords:** Federated learning · Label heterogeneity · Chest x-ray diagnosis · Class prototypes.

## 1 Introduction

Federated learning (FL) [15] enables collaborative training across institutions that cannot share patient data due to regulations such as the GDPR [1]. In practical FL scenarios for chest x-ray diagnosis, institutions annotate different subsets of thoracic abnormalities depending on clinical objectives, patient demographics, and reporting protocols. The resulting diagnostic label spaces are generally partially overlapping or even fully disjoint. A single model cannot be trained over such a non-uniform label space without requiring institutions to disclose their annotations. Most existing FL approaches instead assume a homogeneous label space or require label sharing to construct a global taxonomy. These assumptions are generally impractical in healthcare due to privacy regulations and institutional constraints.

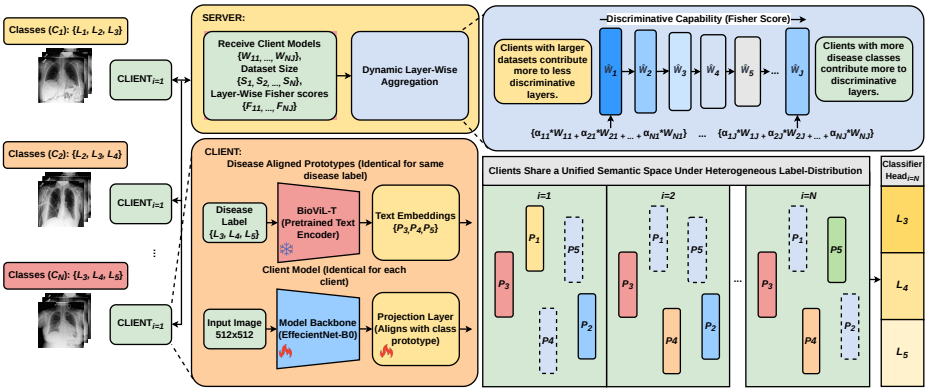



**Fig. 1.** A visual representation of the proposed approach. Each client model consists of common base layers and a projection layer that maps local features to a shared semantic space. Every base layer $j$ of each client $i$ is assigned an aggregation weight $\alpha_{ij}$ that determines its impact on the globally aggregated model.

Personalized FL addresses statistical heterogeneity through methods such as proximal or dynamic regularization [13, 2], clustering [5], multi-task learning [18], transfer learning [21], or meta-learning [8]. Some methods decouple the network into a shared body and a personalized head [3, 14, 7, 6, 16]. Similarly, FD-Fed [11] addresses label heterogeneity without label sharing. However, these methods generally adopt parameter-level personalization and rely on static aggregation, lacking mechanisms to align representations across disjoint label spaces.

To address this challenge, we propose FedDAP (Federated Learning via Disease-Aligned Prototypes), a framework for chest x-ray diagnosis under label heterogeneity that does not require sharing of local labels or raw data (Fig. 1). FedDAP couples shared base layers with a projection layer that maps features into a semantic space defined by disease-aligned class prototypes. These prototypes are derived from a pretrained text encoder with relevant semantic information. The projection layer aligns local feature representations to these prototypes, and a personalized classifier head predicts the disease label from the aligned features. We further propose a dynamic layer-wise aggregation strategy that weights each base layer by its discriminative capability. Clients with larger datasets contribute more to disease-agnostic layers, while clients covering a broader set of disease classes contribute more to discriminative layers. Our contributions are:

- A federated learning framework for chest x-ray diagnosis across clients with heterogeneous label distributions.
- A semantic alignment mechanism that uses disease-aligned prototypes to allow clients to learn semantically consistent feature representations.
- A dynamic layer-wise aggregation strategy that weights each base layer by its discriminative capability.

– Evaluation on three public chest x-ray datasets under 3-client and 6-client configurations.

## 2 Method

We consider $N$ clients, each holding a private dataset $D_i = \{(x, y)\}$ of chest x-ray images with single disease labels drawn from a local label space $C_i$. The label spaces of clients may overlap partially or be entirely disjoint. Aggregation strategies, such as FedAvg [15] assume identical client models trained over a common label space. This makes FedAvg incompatible with our setting. The proposed FedDAP divides each client model into base layers and a classification head. The base layers extract relevant features from the chest x-ray images and the classification head learns to predict the correct disease label using the extracted features. FedDAP introduces a projection layer between the base layers and the classification head of each client model. The projection layer maps feature representations into a shared semantic space, defined by disease-aligned prototypes. These prototypes are derived from a pretrained text encoder with relevant semantic information. Because of this, each disease has a fixed prototype. The objective is to enable clients with heterogeneous label spaces to learn in a shared semantic space defined by disease-aligned prototypes.

### 2.1 Client Training via Disease-Aligned Prototypes

For each disease label $l \in C_i$, we use the pretrained BioViL-T text encoder [4] to encode the corresponding disease name, resulting in a text embedding $\tilde{P}_l$. We then normalize this embedding to define the disease-aligned prototype as $P_l = \tilde{P}_l / \|\tilde{P}_l\|_2$. All clients use the same pretrained text encoder. Because each disease has a fixed prototype, these prototypes create a shared, disease-aligned semantic space for all clients, regardless of their individual label spaces.

For each client $i$, a learnable projection layer $\phi_i(\cdot)$ is placed between the base layers and the classification head of the client model. $\phi_i(\cdot)$ takes the feature representation $h_i(x)$ produced by the base layers for each image $x$ and aligns it to the corresponding disease-aligned prototype $P_l$. We do this by maximizing cosine similarity between the normalized feature representations $v_i(x) = \phi_i(h_i(x))/\|\phi_i(h_i(x))\|_2$ and the corresponding disease-aligned prototype $P_l$. The overall objective for semantic alignment via disease-aligned prototypes is

$$\mathcal{L}_i^{proto} = -\sum_{(x,y)\in D_i} \log \frac{\exp\left(v_i(x)^\top P_y/\tau\right)}{\sum_{l\in C_i} \exp\left(v_i(x)^\top P_l/\tau\right)}, \tag{1}$$

where $\tau$ is a temperature hyperparameter and $P_y$ is the prototype of the ground-truth label. Minimizing (1) clusters the feature representation of a particular disease label around its corresponding disease-aligned prototype and pushes them away from incorrect prototypes. Subsequently, the classifier head $\psi_i(\cdot)$ (which is

an MLP) learns to predict the correct disease label from the aligned representation $v_i(x)$, by minimizing the cross-entropy loss $\mathcal{L}_i^{ce}$. The final client training objective is

$$\mathcal{L}_i = \lambda \mathcal{L}_i^{proto} + (1 - \lambda) \mathcal{L}_i^{ce}, \tag{2}$$

where $\lambda \in [0, 1]$ is a hyperparameter that controls the balance between the prototype loss $\mathcal{L}^{proto}$ and the cross-entropy loss $\mathcal{L}^{ce}$. During training, each client $i$ collects feature representations $h_{ij}(x)$ from data points in each class at every base layer. The client then calculates the intra-class variance $\mathcal{V}_{ij}^{intra}$ (the mean squared Euclidean distance of features to their class centroid) and the inter-class variance $\mathcal{V}_{ij}^{inter}$ (the mean squared Euclidean distance of the class centroids to their global mean) for each base layer $j$. The projection layer is considered one of the $J$ base layers and is also communicated to the server for aggregation.

### 2.2 Server Aggregation

**Weight assignment.** At each communication round, the server receives the base-layer parameters from each client, along with their layer-wise variance statistics $\mathcal{V}_{ij}^{intra}$ and $\mathcal{V}_{ij}^{inter}$. We then use a layer-wise method to aggregate these base layers across all clients. Our aggregation strategy is based on the assumption that highly discriminative layers capture class-specific information, while less discriminative layers (based on the measure of discriminative capability defined in the next paragraph) capture information common to all diseases (disease-agnostic). To reflect this, we give more weight to clients with a wider variety of disease classes when aggregating discriminative layers. For disease-agnostic layers, clients with more training samples get more weight, regardless of how many disease classes they have. This approach helps the aggregated model become more discriminative across different disease classes, while also improving its ability to generalize across a large and diverse set of samples. To achieve this, we assign a specific weight to each base layer of every client model during aggregation.

Let $|D_i|$ denote the number of training samples and $|C_i|$ denote the number of disease classes at client $i$. We normalize these values across clients as $\hat{D}_i = |D_i| / \sum_r |D_r|$ and $\hat{C}_i = |C_i| / \sum_r |C_r|$. The aggregation weight assigned to the $j$-th base layer of client $i$ is computed as

$$\alpha_{ij} = (1 - g_j)\hat{D}_i + g_j \hat{C}_i, \tag{3}$$

where $g_j \in (0, 1)$ is a gating coefficient that controls the balance between data-size weights and class-diversity weights based on the discriminative capability of each $j$-th layer. We quantify the discriminative capability of each $j$-th base layer using the layer-wise variance statistics $\mathcal{V}_{ij}^{intra}$ and $\mathcal{V}_{ij}^{inter}$. Specifically, the server computes the Fisher score [9] for each $j$-th base layer of client $i$ as $F_{ij} = \mathcal{V}_{ij}^{inter} / (\mathcal{V}_{ij}^{intra} + \epsilon)$, where $\epsilon$ is a small constant for numerical stability. We use Fisher scores as it offers a low-cost measure for class separability and can be easily computed for each base layer without compromising clients' data. The

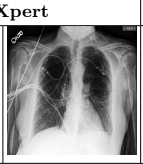
| Dataset | NIH Chest X-ray14 | | CheXpert | | MIMIC-CXR | |
|---|---|---|---|---|---|---|
| Images |  |  |  |  |  |  |
| Ground Truth | Cardiomegaly | Nodule | Pneumothorax | Atelectasis | Pleural Effusion | Pneumonia |
| Detected | Cardiomegaly | Nodule | Pneumothorax | Pneumothorax | Pleural Effusion | Atelectasis |

**Fig. 2.** Classification results of client model 1 using FedDAP on sample images from the NIH Chest X-ray14, CheXpert, and MIMIC-CXR datasets. For each dataset, one correctly and one incorrectly classified example is shown (blue: correct prediction; red: incorrect prediction).

server then averages per-client Fisher scores over the $N$ participating clients as $\bar{F}_j = \frac{1}{N}\sum_{i=1}^{N} F_{ij}$. This gives a common measure of the discriminative capability of each base layer. A higher Fisher score indicates that the feature representations of the corresponding base layer are more separable between classes relative to their spread within a class.

The gating coefficient $g_j$ is obtained using the averaged Fisher score $\bar{F}_j$ as

$$g_j = \frac{\bar{F}_j}{\gamma + \bar{F}_j}, \tag{4}$$

where $\gamma > 0$ is a hyperparameter. Since $g_j$ increases monotonically with $\bar{F}_j$, less discriminative layers ($g_j \to 0$) rely mostly on data-size weights, while highly discriminative layers ($g_j \to 1$) rely mostly on class-diversity weights.

**Layer-wise aggregation.** Let $W_{ij}(t)$ denote the parameters of the $j$-th base layer of client $i$ at communication round $t$. The aggregated parameters of $j$-th base layer are computed as

$$\hat{W}_j(t) = \sum_{i=1}^{N} \alpha_{ij} W_{ij}(t), \tag{5}$$

where $\alpha_{ij}$ is the aggregation weight from Eq. (3). The aggregated base layers $\hat{W}_j(t)$ are broadcast to all clients for the next communication round. Since the Fisher scores are recomputed at every round from the updated layer-wise variance statistics, the aggregation weights adapt dynamically across communication rounds.

## 3 Experiments and Results

**Datasets and Experimental Setup.** We perform experiments on three publicly available chest x-ray datasets: NIH Chest X-ray14 [20], CheXpert [10],

**Table 1.** Classification accuracy (%) and AUROC (%) (mean ± std over 5 runs) for EfficientNetB0 on the NIH Chest X-ray14, CheXpert, and MIMIC-CXR datasets under Protocol 1 (3 clients) and Protocol 2 (6 clients). Best result per column in bold.

| Method | NIH Chest X-ray14 | | CheXpert | | MIMIC-CXR | |
|---|---|---|---|---|---|---|
| | **3 Clients** | **6 Clients** | **3 Clients** | **6 Clients** | **3 Clients** | **6 Clients** |
| | | | *Accuracy (%)* | | | |
| Local | $48.46 \pm 3.92$ | $50.74 \pm 3.35$ | $52.44 \pm 4.27$ | $53.47 \pm 4.58$ | $46.50 \pm 2.70$ | $46.96 \pm 3.83$ |
| FedPer | $46.55 \pm 3.81$ | $50.64 \pm 4.50$ | $45.34 \pm 6.49$ | $51.17 \pm 4.34$ | $41.51 \pm 4.17$ | $43.93 \pm 4.05$ |
| FedPav | $47.53 \pm 3.13$ | $50.91 \pm 4.21$ | $49.77 \pm 4.87$ | $53.84 \pm 4.22$ | $40.30 \pm 4.13$ | $45.98 \pm 2.87$ |
| FedBABU-LI | $48.28 \pm 4.23$ | $50.58 \pm 4.44$ | $53.63 \pm 4.08$ | $52.38 \pm 3.63$ | $44.63 \pm 3.01$ | $47.29 \pm 5.51$ |
| FedRep | $49.07 \pm 2.87$ | $51.44 \pm 2.63$ | $53.15 \pm 3.92$ | $54.97 \pm 4.16$ | $47.21 \pm 1.90$ | $48.18 \pm 3.45$ |
| FD-Fed | $50.68 \pm 3.00$ | $53.13 \pm 3.26$ | $54.33 \pm 3.30$ | $56.40 \pm 4.23$ | $48.46 \pm 2.24$ | $51.37 \pm 2.31$ |
| FedDAP | $\mathbf{51.85 \pm 2.92}$ | $\mathbf{53.97 \pm 3.62}$ | $\mathbf{54.77 \pm 3.38}$ | $\mathbf{57.48 \pm 3.88}$ | $\mathbf{50.27 \pm 3.79}$ | $\mathbf{51.83 \pm 1.60}$ |
| | | | *AUROC (%)* | | | |
| Local | $86.51 \pm 1.58$ | $87.26 \pm 1.46$ | $86.38 \pm 0.74$ | $86.64 \pm 1.61$ | $84.75 \pm 1.55$ | $84.97 \pm 1.25$ |
| FedPer | $86.04 \pm 1.60$ | $87.47 \pm 1.64$ | $82.15 \pm 0.48$ | $85.51 \pm 2.02$ | $82.63 \pm 2.07$ | $83.66 \pm 2.25$ |
| FedPav | $86.19 \pm 1.69$ | $87.63 \pm 1.64$ | $83.90 \pm 1.52$ | $87.29 \pm 1.72$ | $82.10 \pm 3.34$ | $84.59 \pm 1.24$ |
| FedBABU-LI | $86.25 \pm 2.34$ | $86.22 \pm 2.29$ | $86.88 \pm 1.29$ | $86.23 \pm 2.40$ | $81.71 \pm 3.65$ | $84.74 \pm 2.61$ |
| FedRep | $87.25 \pm 1.41$ | $87.93 \pm 1.45$ | $86.79 \pm 0.89$ | $87.51 \pm 1.59$ | $85.52 \pm 1.42$ | $85.78 \pm 0.93$ |
| FD-Fed | $87.45 \pm 1.25$ | $88.35 \pm 1.34$ | $87.05 \pm 0.77$ | $88.07 \pm 1.74$ | $85.94 \pm 1.23$ | $86.70 \pm 0.57$ |
| FedDAP | $\mathbf{87.95 \pm 1.25}$ | $\mathbf{88.76 \pm 1.26}$ | $\mathbf{87.33 \pm 1.04}$ | $\mathbf{88.79 \pm 1.56}$ | $\mathbf{87.54 \pm 1.38}$ | $\mathbf{87.34 \pm 0.73}$ |

and MIMIC-CXR [12]. For each dataset, we exclude multi-label and 'No Finding' samples to obtain a controlled single-label setting. For each dataset, we randomly distribute the labels among clients, so that each client receives data for a distinct set of labels. If multiple clients receive data for the same label, we ensure there is no overlap in the data distributed among those clients. We consider two federated configurations: (i) **Protocol 1**: each dataset is distributed among 3 clients; and (ii) **Protocol 2**: each dataset is distributed among 6 clients. Under both protocols, each client divides its local data into an 80:10:10 ratio for training, validation, and testing. Performance is measured by classification accuracy and the Area Under the Receiver Operating Characteristic Curve (AUROC) for each client on its local test data. In each run, client wise accuracy and AUROC is averaged across all clients. Subsequently, we report the mean and standard deviation over 5 runs.

**Implementation Details.** We compare FedDAP with FedPer [3], FedPav [22], FedRep [7], FD-Fed [11], and FedBABU [16]. We also compare against a 'Local' baseline, in which each client is trained independently without federation. These methods are selected because they are compatible with heterogeneous label spaces. Methods that assume a shared label space across clients, including FedAvg [15], are not applicable to our setting. FedBABU cannot directly handle label heterogeneity, since it initializes a uniform classification head at the server. We instead initialize the classification layer locally for each client and refer to this adapted version as 'FedBABU-LI' . We adopt EfficientNetB0 [19] as the backbone for all methods.

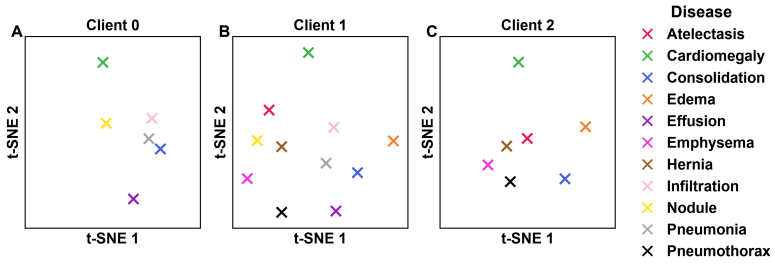



**Fig. 3.** Per-client t-SNE class centroids of disease labels shared across clients for NIH Chest X-ray14 dataset under Protocol 1. We observe that under label heterogeneity, disease-aware semantic alignment produces mostly consistent class centroids across clients.

For each competing method, we use Stochastic Gradient Descent (SGD) [17] with momentum 0.9 as the optimizer. The best learning rate is selected from the range 0.0001 to 0.1. For FedBABU-LI and FD-Fed, we select a second learning rate for the fine-tuning steps. A single fine-tuning step is performed per round for FedRep and FD-Fed. FedBABU-LI instead performs 5 fine-tuning steps after the final communication round. FedPer and FedPav do not require fine-tuning. For FedDAP, the adapt phase fine-tunes the personalized classifier head for a single step after each round. The loss weight $\lambda$ is fixed to 0.5, and the temperature hyperparameter $\tau$ is selected from the range 0.01 to 0.1. All hyperparameters are tuned using validation performance. We run each experiment for 50 communication rounds, with 5 local training steps per round.

**Comparative Analysis.** Each experiment is repeated for five runs under both the 3-client and 6-client configurations. Results are reported in Table 1. FedDAP achieves the highest accuracy and AUROC across all three datasets and both client settings. The largest gain over the strongest baseline (FD-Fed) occurs on MIMIC-CXR under Protocol 1, where accuracy improves from 48.46% to 50.27% (a relative improvement of 3.7%) and AUROC from 85.94% to 87.54%. Methods that don't perform classifier fine-tuning degrade under label heterogeneity (FedPer and FedPav), falling below the Local baseline on MIMIC-CXR under Protocol 1 (41.51% and 40.30% accuracy versus 46.50%). FedDAP exceeds the Local baseline under both protocols on each dataset, with the largest gains on MIMIC-CXR. We observe that accuracy is consistently low despite the high AUROC. This is because accuracy requires the correct disease label to have the highest predicted probability, whereas AUROC remains high as long as the correct disease receives a high score, even when it is not the top prediction. Qualitative results in Fig. 2 show the classification outputs of client model 1 on two randomly selected test samples per dataset.

**Cross-Client Semantic Space Alignment.** To assess whether clients learn consistent representations for disease labels shared across clients under label heterogeneity, we pool the feature representations of each client's local data for

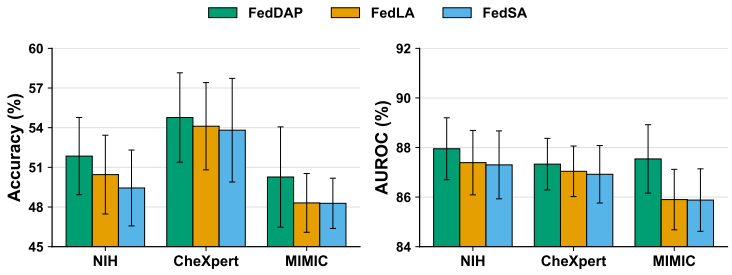



**Fig. 4.** Component-wise ablation study of FedDAP: FedLA removes disease-aware semantic alignment while keeping layer-wise aggregation. Further, FedSA replaces layer-wise aggregation with standard FedAvg. Results show accuracy (%) and AUROC (%) (mean ± std over 5 runs) on NIH Chest x-ray14, CheXpert, and MIMIC-CXR under Protocol 1 (3 clients).

the NIH Chest X-ray14 dataset and fit a single t-SNE, ensuring a common embedding space and avoiding misalignment from separate projections. We then plot the per-client class centroids for shared disease labels in the common space (Fig. 3). Despite differences in their overall label spaces, the centroids of shared diseases remain largely aligned across clients. This demonstrates that anchoring features to disease-aligned prototypes maps semantically equivalent disease representations into a shared feature space without sharing local labels.

### 3.1 Ablation Studies

We evaluate the significance of each component of FedDAP by systematically removing each component and measuring the drop in classification accuracy and AUROC (Fig. 4). FedDAP denotes our full proposed framework, FedLA denotes our method without the disease-aware semantic alignment, and FedSA replaces the layer-wise aggregation with standard FedAvg. FedDAP achieves accuracies of 51.85/54.77/50.27% and AUROC scores of 87.95/87.33/87.54% on NIH Chest X-ray14, CheXpert, and MIMIC-CXR, respectively. Removing disease-aware semantic alignment (FedLA) lowers accuracy to 50.45/54.11/48.31% and AUROC to 87.39/87.04/85.90%. Further, FedSA reduces accuracy to 49.44/53.81/48.28% and AUROC to 87.30/86.92/85.88%. The consistent drop in performance across all three datasets confirms that both the dynamic layer-wise aggregation and the disease-aware semantic alignment contribute meaningfully to the overall effectiveness of FedDAP.

## 4 Conclusion

We introduce FedDAP, a federated learning framework for chest x-ray diagnosis under label heterogeneity. By aligning features to disease-aligned prototypes derived from a pretrained BioViL-T text encoder, clients with disjoint label

spaces are trained jointly in a shared semantic space without sharing local labels or raw data. A dynamic layer-wise aggregation strategy weights each base layer by its discriminative capability using inter-class and intra-class feature statistics. Experiments across three public chest x-ray datasets and two client configurations demonstrate consistent improvements over existing personalized federated learning methods. Ablation experiments validate the significance of semantic alignment via disease-aligned prototypes. Our present study adopts a single-label protocol, with each dataset split internally across clients. Future work will extend the prototype-based training and layer-wise aggregation to the multi-label and multi-institutional settings. We will also evaluate alternative biomedical and general-purpose text encoders and study robustness to ambiguous or institution-specific terminology. We also plan to extend the framework to model heterogeneity, multi-modal medical data, and task-heterogeneous federated learning settings.

## References


1. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation) (2016), https://data.europa.eu/eli/reg/2016/679/oj, oJ L 119, 4.5.2016, p. 1–88
2. Acar, D.A.E., Zhao, Y., Matas, R., Mattina, M., Whatmough, P., Saligrama, V.: Federated learning based on dynamic regularization. In: International Conference on Learning Representations (ICLR) (2021)
3. Arivazhagan, M.G., Aggarwal, V., Singh, A.K., Choudhary, S.: Federated learning with personalization layers. arXiv preprint arXiv:1912.00818 (2019)
4. Bannur, S., Hyland, S., Liu, Q., Pérez-García, F., Ilse, M., Castro, D.C., Boecking, B., Sharma, H., Bouzid, K., Thieme, A., Schwaighofer, A., Wetscherek, M., Lungren, M.P., Nori, A., Alvarez-Valle, J., Oktay, O.: Learning to exploit temporal structure for biomedical vision-language processing. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15016–15027 (2023). https://doi.org/10.1109/CVPR52729.2023.01442
5. Briggs, C., Fan, Z., Andras, P.: Federated learning with hierarchical clustering of local updates to improve training on non-iid data. In: 2020 International Joint Conference on Neural Networks (IJCNN). pp. 1–9. IEEE (2020)
6. Chen, H.Y., Chao, W.L.: On bridging generic and personalized federated learning for image classification. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=I1hQbx10Kxn
7. Collins, L., Hassani, H., Mokhtari, A., Shakkottai, S.: Exploiting shared representations for personalized federated learning. In: International Conference on Machine Learning. pp. 2089–2099. PMLR (2021)
8. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: International Conference on Machine Learning. pp. 1126–1135. PMLR (2017)
9. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Annals of Eugenics **7**(2), 179–188 (1936)

10. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D.A., Halabi, S.S., Sandberg, J.K., Jones, R., Larson, D.B., Langlotz, C.P., Patel, B.N., Lungren, M.P., Ng, A.Y.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison (2019), https://arxiv.org/abs/1901.07031
11. Jamir, O., Paul, A.: Feature-driven layer specialization for label heterogeneous federated learning. Neurocomputing **670**, 132620 (2026). https://doi.org/https://doi.org/10.1016/j.neucom.2026.132620, https://www.sciencedirect.com/science/article/pii/S0925231226000172
12. Johnson, A.E.W., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data **6**(1), 317 (2019). https://doi.org/10.1038/s41597-019-0322-0, https://doi.org/10.1038/s41597-019-0322-0
13. Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V.: Federated optimization in heterogeneous networks. In: Dhillon, I., Papailiopoulos, D., Sze, V. (eds.) Proceedings of Machine Learning and Systems. vol. 2, pp. 429–450 (2020)
14. Liang, P.P., Liu, T., Ziyin, L., Allen, N.B., Auerbach, R.P., Brent, D., Salakhutdinov, R., Morency, L.P.: Think locally, act globally: Federated learning with local and global representations. arXiv preprint arXiv:2001.01523 (2020)
15. McMahan, H.B., Moore, E., Ramage, D., Hampson, S., et al.: Communication-efficient learning of deep networks from decentralized data. In: Proc. 20th Int'l Conf. Artificial Intelligence and Statistics (AISTATS). pp. 1273–1282 (2017)
16. Oh, J., Kim, S., Yun, S.Y.: FedBABU: Toward enhanced representation for federated image classification. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=HuaYQfggn5u
17. Robbins, H., Monro, S.: A stochastic approximation method. The Annals of Mathematical Statistics **22**(3), 400–407 (1951). https://doi.org/10.1214/aoms/1177729586, https://www.jstor.org/stable/2236626
18. Smith, V., Chiang, C.K., Sanjabi, M., Talwalkar, A.: Federated multi-task learning. arXiv preprint arXiv:1705.10467 (2017)
19. Tan, M., Le, Q.V.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning (ICML). pp. 6105–6114 (2019), https://arxiv.org/abs/1905.11946
20. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. Proceedings of the IEEE conference on computer vision and pattern recognition pp. 2097–2106 (2017), https://arxiv.org/abs/1705.02315
21. Yang, H., He, H., Zhang, W., Cao, X.: Fedsteg: A federated transfer learning framework for secure image steganalysis. IEEE Transactions on Network Science and Engineering **8**(2), 1084–1094 (2021). https://doi.org/10.1109/TNSE.2020.2996612
22. Zhuang, W., Wen, Y., Zhang, X., Gan, X., Yin, D., Zhou, D., Zhang, S., Yi, S.: Performance optimization of federated person re-identification via benchmark analysis. In: Proceedings of the 28th ACM International Conference on Multimedia. p. 955–963. MM '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3394171.3413814, https://doi.org/10.1145/3394171.3413814