PICSR: Prototype-Informed Cross-Silo Router for Federated Learning

1st Eric Enouen
Computer Science & Engineering
The Ohio State University
enouen.9@osu.edu

2nd Sebastian Caldas
Auton Lab
Carnegie Mellon University
scaldas@cmu.edu

3rd Artur Dubrawski
Auton Lab
Carnegie Mellon University
awd@cs.cmu.edu

***Abstract*—Recent developments in federated learning (FL) have focused on pushing forward along the axis of predictive performance, maximizing metrics such as accuracy under a wide variety of settings. However, little work has been done to explain the knowledge differences between institutions and the benefits of collaboration, which is critical in cross-silo federated learning domains, e.g., healthcare or banking. We aim to develop further along this axis of interpretability, introducing Prototype-Informed Cross-Silo Router (PICSR[1]). We allow each silo to train their own model and utilize a mixture of experts approach to explicitly score each expert model in accordance to its ability to perform well on a new incoming sample. By embedding the decisions in prototypes from each silo, we are able to ground the predictions in the underlying dataset's distributions. We evaluate our approach on a Heart Disease dataset to show our approach has strong performance, in addition to a new form of interpretability, compared to current federated approaches.**

I. Introduction

In practice, federated learning (FL) frequently experiences variation among clients, e.g., variation in data space, statistical heterogeneity, and systems heterogeneity [1]. Specifically in the cross-silo setting, institutions typically have large datasets from different underlying distributions and so each have a different knowledge base. Thus, it is critical that proposed methods are capable of leveraging this heterogeneity in order to build better models.

For example, in healthcare settings, the patient population of a given hospital is primarily based on its location, which is correlated to ethnicity and socioeconomic status [2]. When a new patient is not represented in the population distribution of a hospital, a model's performance on this patient can worsen drastically [3, 4].

This bias is detrimental to the use of AI in the healthcare field, but heterogeneous federated learning approaches can learn and utilize the differences amongst institutions, training a model that performs better despite the heterogeneity [1]. However, the vast majority of these methods focus on performance, resulting in a final black-box model that must be trusted blindly. This is clearly insufficient in real-world domains, where stakeholders need to understand how institutions differ from their own, what benefits they are receiving from participation, and the drawbacks of their own model [5].

[1]pronounced as 'Pixar'.

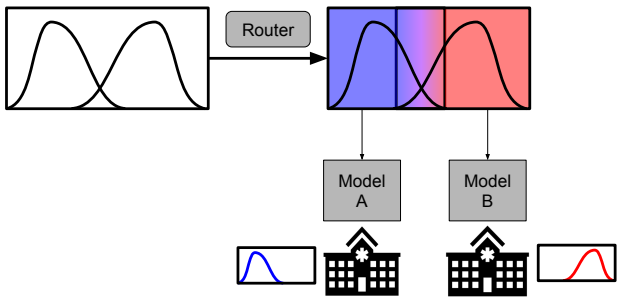

Fig. 1. Illustration of router's goal. Each institution has their own underlying dataset distributions, and the router needs to map samples to the model that is best trained to perform well on them by weighting their predictions.

There have been some recent approaches developing interpretable federated learning. [6] proposes a solution for more interpretable client selection through an uncertainty estimation technique, and [7, 8] present many more interpretable methods for federated learning such as post-hoc explanations methods and interpretable-by-design federated models, as well as interpretable sample and feature selection. A general framework was also proposed to analyze the level of trustworthiness of federated learning models [9]. However, these works do not attempt to explain the differences between models and institutional datasets.

Different from these approaches, we aim to answer the following question: **how do other institutions differ from my own?** To accomplish this we propose a novel Prototype-Informed Cross-Silo Router (PICSR) to learn to map from samples to silos. We utilize a mixture of experts framework to enable each expert to specialize to its own institution's data while the router decides how to ensemble their predictions. This design allows for not only high performance in the Cross-Silo setting, but the explanations provided by the model allow for stakeholders to analyze the decisions made.

Our method is a simple approach that can explicitly make use of the heterogeneity between participants to build a model that can improve the performance for all silos.

We make the following contributions:

- We propose a novel cross-silo framework for combining the capabilities of individual models using a mixture of

experts approach.
- We embed the predictions of our router in the prototypes from each silo, creating a more explainable pipeline for stakeholders.
- We analyze our method on a real-life healthcare dataset, showing both the performance and interpretability aspects of our approach.

II. Related Work

A. Mixture of Experts

In the Mixture of Experts (MoE) framework, we aim to divide the problem space into different partitions based on a gating function, where one expert model is responsible for each partition [10]. This approach simplifies the problem space into easier subtasks that individual models can specialize to. In federated learning, MoE has been applied for crowdsourced training of deep networks [11], utilizing a decentralized MoE to efficiently select the appropriate experts for training a large language model. However, this approach focuses on performance/efficiency, and does not look at interpretability at all.

MoE has also been used in personalized federated learning to balance between the generalization of a global model and the specialization of a local model [12, 13, 14]. These works use local gating functions to personalize the resulting model to each client. [15] also learns local gating functions, but instead trains K global experts and allows each client to select which ones to use. FedEM [16] assumes each client's data follows a mixture of M underlying distributions and learns M shared components. FedMN [17] allows each client to select a combination of modular blocks with a routing hypernetwork. Most similarily to us, [18, 19] let each institution train its own model and learn an ensembling approach to combine them. However, these approaches learn sample-agnostic methods that cannot be used to explain the differences among institutions

In our work, we look specifically at the cross-silo setting where each participant is assumed to have enough data to train its own model. Different from prior work on MoE in federated learning, we select our experts to be the local model from each silo, and train one global routing model to learn how to combine their predictions. By inspecting our interpretable router, we enable stakeholders to analyze the capabilities of their model compared to others, as well as how the sample distribution varies amongst silos.

B. Prototypes

Prototypes are generally assumed to be some data point that is more generally representative of the underlying data distribution. Additionally, MoE is a useful tool for improving performance, but can also be used to gain insight about how the problem decomposes amongst the different experts [20, 21]. The use of prototypes for federated learning is not new [22, 23], but we look to utilize prototypes for the purpose of interpretability, not just performance. We embed these prototypes in the router in a similar fashion to [24, 25]. We create a routing model based on the ProtoPNet architecture [26] where the final prediction of the model is grounded in the similarity scores to the prototypical samples from each silo.

C. Data Heterogeneity

Heterogeneity is common in federated learning [1] and numerous techniques have been proposed to tackle data heterogeneity specifically. For example, one recent work on heterogeneous FL uses unlabeled public data and self-supervised techniques to ensure logit similarity between models [27]. Meanwhile, another work uses an alignment step to ensure all clients are sharing centered information [28].

Instead, in this work we focus directly on the routing model to uncover more information about how different silos may differ. We first build a general cross-silo federated learning pipeline that utilizes prototypes to make predictions, and then we focus on extracting explanations from the prototypical layer. The focus on the routing model's ability to give an explanation along with its prediction enables stakeholders to understand why a given sample would benefit from being routed to a different local expert. This is of particular importance in the healthcare setting, where clinicians want to understand the differences in expertise between centers and contrast these against their own intuition.

III. Methodology

A. Background

We assume there are K participants, and that each participant i has a local dataset $D_i = \{(X_i, Y_i)|X_i \in R^{N_i \times D}, Y_i \in R^{N_i \times C}\}$ where N_i is the number of samples, D is the number of features, and C is the number of classes. Each participant aims to minimize the loss of its local objective function $F(E_i(w_i))$ given some local model $E_i(w_i)$.

Further, following [27], we highlight the data heterogeneity problem as below.

$$P_i(X) \neq P_j(X) \tag{1}$$

where $P_i(X), P_j(X)$ are distributions of the features in the local dataset i, j. Each participant experiences some domain shift even though the labels are the same.

B. Model Architecture

Our framework is depicted in Figure 3. We utilize the ProtoPNet [26] architecture to train a routing model to ensemble a group of expert models together in an explainable manner. There are two core components to our approach: our federated mixture of experts approach to learning an appropriate weighted ensemble for each sample, and the prototype embedding in the router to ensure grounded and explainable decisions.

1) Federated Mixture of Experts: We assume the datasets to remain local and to experience domain shift. Thus, we can treat each local model as an expert on the distribution that its training dataset is sampled from. In the extreme case, these datasets are sampled from entirely disjoint distributions, meaning that each local model trained on its own would

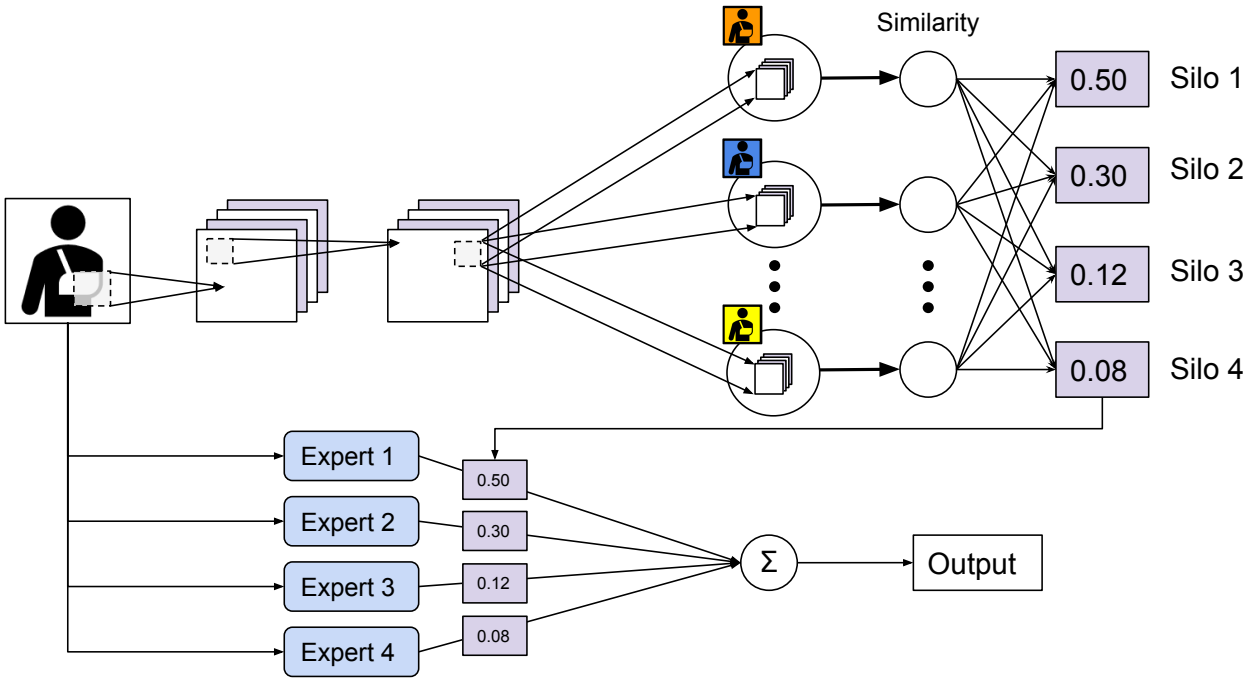

Fig. 2. Overview of core algorithm. Router is built on the ProtoPNet architecture and outputs the weights to ensemble expert predictions together.

perform optimally. However, this is not the case in the real world, as the distributions overlap and the local models can benefit from sharing information with one another.

To account for this, we train a router to output a weight for each expert: $h(x;\Theta) \in R^K$ given a sample x and the network parameters Θ and then perform soft-routing as shown in the equation below:

$$\hat{y} = \sum_{i=1}^{K} h(x;\Theta)_i \times E_i(x, w_i) \tag{2}$$

This allows each expert to learn to specialize to the distribution that they are most equipped to handle, and allows the router to divide up the problem space.

2) Prototype-Informed Router: By allowing the router to learn how to divide the problem space, stakeholders can analyze the outputs of the router to understand how their local model differs from other participants. To further improve this explanation, we implement an architecture similar to ProtoPNet [26] to ground the predictions of the router in prototypical samples. These prototypical samples in our case are defined as the average sample for all of the training samples in each participant's dataset. However, these prototypes could also be learned in the future to more closely follow ProtoPNet's architecture.

As shown in Figure 2, we begin with some sample X. We then utilize some embedding function $f(x)$ to convert our sample into its latent vector z. This latent vector can then be compared to each prototypical latent vector using the L_2 distance as shown below:

$$dist(z, p_f) = ||z - p_f||_2 \tag{3}$$

where p_f indexes each prototypical latent vector. This outputs F distance scores where F is the number of prototypes. Then, the similarity is computed by inverting the distance: $\frac{1}{1+dist(z,p_f)}$. These similarity scores are then fed into a final linear prediction layer to compute the importance of each prototype for the final prediction. The output of the routing model is then utilized to weight the expert models and compute the final prediction.

C. Training Process

We train this model following a typical federated framework. The router is passed at each communication step to each participant, as well as a frozen copy of each expert model. Then each participant minimizes their objective function given the weighted ensemble of experts as a model. This equates to solving the following optimization problem.

$$min \sum_{i=1}^{N} CrossEntropy(\hat{y}, y) \tag{4}$$

where $\hat{y}$ is defined in equation 2.

Each participant can train the copy of the global router as well as their own local model. The updates are then averaged using FedAvg to create the next router θ_{d+1}

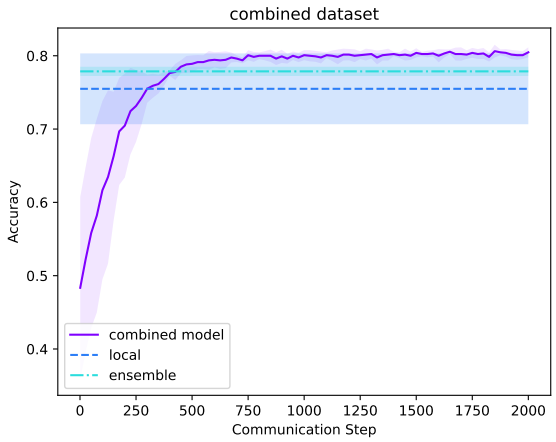

Fig. 3. Accuracy of our approach as you increase the number of communication steps.

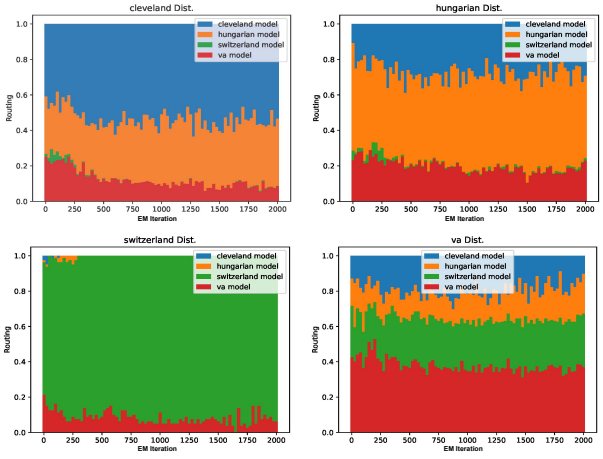

Fig. 4. Routing Distribution for all four silos in the Heart Disease dataset

IV. Case Study: Heart Disease

1) Dataset: We first evaluate our approach on the Fed-Heart-Disease dataset, following the implementation of [29]. The dataset has 740 samples naturally split up into four silos based on the hospital: Cleveland (303), Hungarian (261), Switzerland (46), and VA Long Beach (130). Each sample has 13 features, and the goal is to predict whether a patient has heart disease or not.

2) Methods: For this dataset, we initialize one prototype per silo by computing the average patient's feature vector, and then train the router to ground its predictions in these prototypes. Each expert model is a Logistic Regression model and the router is a fully connected network with one hidden layer. The embedded space is 20 dimensions. We train for 2000 communication steps with one local epoch per communication step, to ensure that all models converge and we report the average final accuracy over five runs.

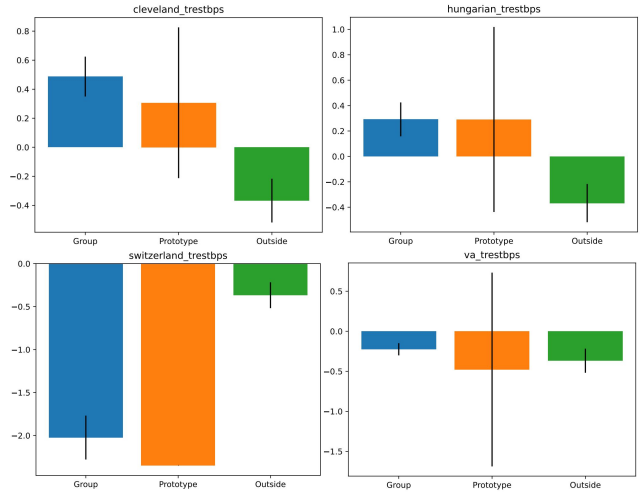

Fig. 5. Analysis of the resting blood pressure of patients. 'Group' is the average feature for all of the samples routed to the silo's model, 'Prototype' is the average feature for training samples from the given silo, and 'Outside' averages over all samples not routed to the silo's model.

A. Results

We compare our method to two core baselines: Local and Ensemble. Local is characterized by local models trained only on their local datasets, and then tested on their local testing datasets. Ensemble outputs the same weighting for each expert, and averages their outputs for the final prediction.

1) PICSR Improves Performance: As shown in table I, PICSR consistently improves the performance compared to the Local baseline, except for Switzerland's test set where it slightly degrades performance. Additionally, PICSR also reliably outperforms an ensemble approach except on Cleveland's test set, but the two means are not statistically significantly different. This results in a large overall boost by using our method of 2.6% over a naive ensemble and 5.88% over local models when all of the test datasets are combined together. This is also visualized over the total number of communication steps in Figure 3. Hence, we can conclude that PICSR is able to improve predictive performance over these baselines on the Heart Disease dataset.

2) PICSR Routes Reliably: In Figure 4, we show that our router learns a function that both performs well and lines up with our intuition. Each local dataset is primarily mapped to the silo that was trained on it, but there are also some key areas where information can be shared across silos. For example, the Cleveland and Hungarian models both improved each other, and this is reflected in the routing distribution for these two datasets. However, the local model for Switzerland is most performant on its own test dataset and so almost all samples are routed back to it. We also show that our approach does not collapse to any one model, and instead learns to map each sample to the model that can perform best on it.

3) Prototypes Ground the Router: By analyzing Figure 5, we can see that the router learns to map samples according to the embedded prototypes. In the Switzerland test dataset,

	Cleveland	Hungary	Switzerland	VA	Total
Local	73.65±0.47	75.51±0.45	**93.75±0.00**	73.33±1.41	74.59±4.83
Ensemble	**78.46±1.44**	77.08±0.55	85.00±3.06	75.56±0.00	77.87±0.63
Ours	77.88±2.11	**80.67±1.31**	90.00±3.06	**82.67±1.66**	**80.47±0.69**

TABLE I

FINAL MEAN ACCURACY OVER FIVE RUNS FOR LOCAL, ENSEMBLE, AND OUR APPROACH. WE SHOW A LARGE BOOST IN THE PERFORMANCE OF TWO SILOS, AND CLOSE PERFORMANCE IN THE OTHER TWO SILOS. WE SHOW AN AVERAGE PERFORMANCE BOOST OF 2.6% OVER A NAIVE ENSEMBLE.

all of their training patients have a very low resting blood pressure, and the majority of samples routed to Switzerland share that low blood pressure. Additionally, both Cleveland and Hungary's training datasets have a much higher resting blood pressure compared to Switzerland, and this trend is also continued in the samples that are routed to Cleveland and Hungary's models.

V. CONCLUSION

We have proposed a new method, embedding prototypes within a mixture of experts formulation to intelligently aggregate local models. This approach is not only more performant than our baselines, but the router's prediction can also enable stakeholders to better understand the differences across institutions.

VI. ACKNOWLEDGEMENTS

This work was partially supported by the Research Experiences for Undergraduates program under National Science Foundation grant 1730147. The author would also like to thank Rachel Burcin, John Dolan, and the RISS staff for organizing this program, the Auton Lab for its support, and their labmates for the fruitful discussions.

REFERENCES

[1] D. Gao, X. Yao, and Q. Yang, "A survey on heterogeneous federated learning," *arXiv preprint arXiv:2210.04505*, 2022.

[2] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings *et al.*, "Advances and open problems in federated learning," *Foundations and Trends® in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[3] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on fairness, accountability and transparency*. PMLR, 2018, pp. 77–91.

[4] A. R. Martin, M. Kanai, Y. Kamatani, Y. Okada, B. M. Neale, and M. J. Daly, "Current clinical use of polygenic scores will risk exacerbating health disparities," *Nature genetics*, vol. 51, no. 4, p. 584, 2019.

[5] S. Caldas, J. H. Yoon, M. R. Pinsky, G. Clermont, and A. Dubrawski, "Understanding clinical collaborations through federated classifier selection," in *Machine Learning for Healthcare Conference*. PMLR, 2021, pp. 126–145.

[6] Z. Qin, L. Yang, Q. Wang, Y. Han, and Q. Hu, "Reliable and interpretable personalized federated learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 20 422–20 431.

[7] J. L. C. Bárcena, M. Daole, P. Ducange, F. Marcelloni, A. Renda, F. Ruffini, and A. Schiavo, "Fed-xai: Federated learning of explainable artificial intelligence models," 2022.

[8] A. Li, R. Liu, M. Hu, L. A. Tuan, and H. Yu, "Towards interpretable federated learning," *arXiv preprint arXiv:2302.13473*, 2023.

[9] P. M. S. Sánchez, A. H. Celdrán, N. Xie, G. Bovet, G. M. Pérez, and B. Stiller, "Federatedtrust: A solution for trustworthy federated learning," *arXiv preprint arXiv:2302.09844*, 2023.

[10] S. Masoudnia and R. Ebrahimpour, "Mixture of experts: a literature survey," *The Artificial Intelligence Review*, vol. 42, no. 2, p. 275, 2014.

[11] M. Ryabinin and A. Gusev, "Towards crowdsourced training of large neural networks using decentralized mixture-of-experts," *Advances in Neural Information Processing Systems*, vol. 33, pp. 3659–3672, 2020.

[12] B. Guo, Y. Mei, D. Xiao, and W. Wu, "Pfl-moe: personalized federated learning based on mixture of experts," in *Web and Big Data: 5th International Joint Conference, APWeb-WAIM 2021, Guangzhou, China, August 23–25, 2021, Proceedings, Part I 5*. Springer, 2021, pp. 480–486.

[13] F. Hanzely and P. Richtárik, "Federated learning of a mixture of global and local models," *arXiv preprint arXiv:2002.05516*, 2020.

[14] E. L. Zec, O. Mogren, J. Martinsson, L. R. Sütfeld, and D. Gillblad, "Specialized federated learning using a mixture of experts," *arXiv preprint arXiv:2010.02056*, 2020.

[15] M. Reisser, C. Louizos, E. Gavves, and M. Welling, "Federated mixture of experts," *arXiv preprint arXiv:2107.06724*, 2021.

[16] O. Marfoq, G. Neglia, A. Bellet, L. Kameni, and R. Vidal, "Federated multi-task learning under a mixture of distributions," *Advances in Neural Information Processing Systems*, vol. 34, pp. 15 434–15 447, 2021.

[17] T. Wang, W. Cheng, D. Luo, W. Yu, J. Ni, L. Tong, H. Chen, and X. Zhang, "Personalized federated learning via heterogeneous modular networks," in *2022 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2022, pp. 1197–1202.

[18] J. M. Reps, R. D. Williams, M. J. Schuemie, P. B. Ryan, and P. R. Rijnbeek, "Learning patient-level prediction models across multiple healthcare databases: evaluation

of ensembles for increasing model transportability," *BMC medical informatics and decision making*, vol. 22, no. 1, p. 142, 2022.

[19] J. Luo and S. Wu, "Adapt to adaptation: Learning personalization for cross-silo federated learning," *arXiv preprint arXiv:2110.08394*, 2021.

[20] Y. Krishnamurthy and C. Watkins, "Interpretability in gated modular neural networks," in *eXplainable AI approaches for debugging and diagnosis.*, 2021.

[21] A. Chaoub, C. Cerisara, A. Voisin, and B. Iung, "Towards interpreting deep learning models for industry 4.0 with gated mixture of experts," in *2022 30th European Signal Processing Conference (EUSIPCO)*. IEEE, 2022, pp. 1412–1416.

[22] Y. Tan, G. Long, L. Liu, T. Zhou, Q. Lu, J. Jiang, and C. Zhang, "Fedproto: Federated prototype learning across heterogeneous clients," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 8, 2022, pp. 8432–8440.

[23] Y. Dai, Z. Chen, J. Li, S. Heinecke, L. Sun, and R. Xu, "Tackling data heterogeneity in federated learning with class prototypes," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 6, 2023, pp. 7314–7322.

[24] G. Armano and N. Hatami, "Mixture of random prototype-based local experts," in *Hybrid Artificial Intelligence Systems: 5th International Conference, HAIS 2010, San Sebastián, Spain, June 23-25, 2010. Proceedings, Part I 5*. Springer, 2010, pp. 548–556.

[25] ——, "An improved mixture of experts model: Divide and conquer using random prototypes," in *Ensembles in Machine Learning Applications*. Springer, 2011, pp. 217–231.

[26] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su, "This looks like that: deep learning for interpretable image recognition," *Advances in neural information processing systems*, vol. 32, 2019.

[27] W. Huang, M. Ye, and B. Du, "Learn from others and be yourself in heterogeneous federated learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 143–10 153.

[28] M. Mendieta, T. Yang, P. Wang, M. Lee, Z. Ding, and C. Chen, "Local learning matters: Rethinking data heterogeneity in federated learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8397–8406.

[29] J. Ogier du Terrail, S.-S. Ayed, E. Cyffers, F. Grimberg, C. He, R. Loeb, P. Mangold, T. Marchand, O. Marfoq, E. Mushtaq, B. Muzellec, C. Philippenko, S. Silva, M. Teleńczuk, S. Albarqouni, S. Avestimehr, A. Bellet, A. Dieuleveut, M. Jaggi, S. P. Karimireddy, M. Lorenzi, G. Neglia, M. Tommasi, and M. Andreux, "Flamby: Datasets and benchmarks for cross-silo federated learning in realistic healthcare settings," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 5315–5334.