Uncertainty Quantification in Federated Learning for Heterogeneous Health Data

Anonymous Author(s)*

ABSTRACT

Machine learning-boosted safety-critical healthcare applications require the learned models to be both high-performing and uncertainty-aware, yet this is challenging due to insufficient data volume at the sources (e.g., in hospitals). Federated learning (FL) enables such data sources to learn in collaboration without transferring their sensitive data to a mutually trusted server, overcoming the barrier between data aggregation and model performance. Since FL is achieved by globally synchronizing the models learned locally, the heterogeneity in local health data, arising from variations in technologies, patient demographics, and disease prevalence, presents significant challenges to FL and correspondingly to uncertainty quantification. It is unclear how reliable the uncertainty is to infer the confidence of the diagnoses made by an FL model.

In this paper, we present the first evaluation of the quality of uncertainty in realistic healthcare FL settings. Our experiments on real-world applications cover tabular data-based heart disease prediction, image-driven skin pathology screening, and physiological signal-based activity detection tasks. Three uncertainty quantification methods that were previously proposed in standard centralized deep learning are adapted to a variety of FL algorithms for comparison. We found that federated deep ensembles perform consistently better than other federated uncertainty quantification methods, and personalization, i.e., training collaboratively but remaining customized models, can further enhance the performance (with an improvement by up to 19%). Our work paves the way for the future development of federated uncertainty quantification approaches.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; **Uncertainty quantification**; **Distributed artificial intelligence**; • **Applied computing** → **Health informatics**.

KEYWORDS

Uncertainty quantification, federated learning, healthcare, data heterogeneity, personalization

1 INTRODUCTION

With the proliferation of clinical data and devices, deep learning is being increasingly applied in the medical field, proving its effectiveness in various applications [6, 9, 25]. For this research to transition into clinical practice, the safety-critical nature of healthcare applications necessitates the models' ability to quantify uncertainty [12]. While deep learning models are known to be overconfident [8], a trustworthy health model needs to convey uncertainty when its prediction is likely to be wrong, prompting human intervention and better risk management.

Nevertheless, overfitting caused by insufficient data poses a significant challenge for individual parties to train high-performing deep learning models that are able to accurately estimate uncertainty. With limited data samples, these parties need to collaborate in model training, but they face obstacles in directly sharing data due to privacy regulations and the risk of data misuse. Federated learning (FL) offers a promising solution where multiple clients, such as hospitals and user devices, to collaboratively train a model without sharing their private data [10]. In FL, the parameters of the local models are aggregated after each training round to learn a global federated model from the data of all clients. The most commonly used method is called FedAvg, which employs weighted averaging based on the proportional data size of each client [19].

Commonly, data collected from different hospitals or health monitoring devices vary in the technologies used, patient demographics, and disease prevalence. This complicated data heterogeneity poses a huge challenge to learning a single global model performing well on all clients. Personalized federated learning (pFL), which involves learning a tailored model for each client, has emerged as a promising approach to address this challenge [21]. Previous studies [18] have demonstrated the effectiveness of personalization in FL for healthcare. However, the behavior of uncertainty estimation in FL in the presence of heterogeneity and the effectiveness of personalization is still unknown and requires further exploration and research.

This work fills this gap by evaluating different uncertainty methods combined with various FL and pFL strategies in real-world heterogeneous healthcare datasets. In particular, we aim to address the following research questions:

- **RQ1**: How does the standard FL (FedAvg) perform in uncertainty quantification with heterogeneous health data?
- **RQ2**: Can personalization of the federated models effectively improve uncertainty quantification?
- **RQ3**: Which uncertainty quantification method is more effective in FL with heterogeneous health data?

With these research questions in mind, we built an uncertainty estimation framework on several real-world multi-site healthcare datasets. Our framework incorporates Monte Carlo (MC) Dropout [7] and deep ensembles [13] as the uncertainty methods. We apply them to three healthcare applications: heart disease detection, human activity recognition, and melanoma class prediction. The datasets used encompass different types of input data, including images, time series, and pre-extracted features. On each dataset and application, the effectiveness of existing uncertainty qualification methods is measured across multiple FL strategies of different degrees of personalization.

To the best of our knowledge, this work is the first to delve into the unique challenges brought by FL data heterogeneity on uncertainty estimation. This is also the first work studying the role of personalized FL on uncertainty quantification. From extensive exploratory experiments, our main conclusions are as follows:

- Compared to centralized learning (where a model is learned on the data aggregated across all the clients) the quality of uncertainty degrades in FL with heterogeneous data along with the classification performance.
- Personalized federated models not only increase accuracy but also enhance uncertainty estimation, by up to 12% in misclassification detection and 19% in selective prediction. This is even better than performance in centralized setting.
- Deep ensembles perform consistently well among the uncertainty quantification methods in personalized FL, with an advantage of up to 4% in misclassification detection, in addition to its accuracy gain.

Our study paves the way for trustworthy deep learning for healthcare by incorporating uncertainty quantification and federated learning in one framework and identifies personalized FL as a promising approach addressing data heterogeneity for both classification and uncertainty estimation.

2 PRELIMINARIES AND RELATED WORK

2.1 Federated Learning

In federated learning, the standard objective is to learn a global model θ that performs well for all the clients on average. The objective is formulated by Li te al. [14] as

$$\min_{\theta} F(\theta) := \sum_{k=1}^{K} p_k \cdot F_k(\theta), \tag{1}$$

where K is the number of participating clients, $F_k(\theta)$ is the local objective function for the k-th client weighted by p_k ($p_k \geq 0$ and $\sum_k p_k = 1$). The federated training is done by iterative local training and global synchronization . At the start of each round, the server sends the global model to participating clients. Each client independently trains the model using its local dataset in parallel and sent it back to the server for aggregation. The aggregation function varies across strategies.

FedAvg [19], the most commonly used strategy, takes a weighted average proportional to the number of samples of each client, i.e., $p_k = \frac{n_k}{n}$. Despite being proven to be empirically effective, the convergence of FedAvg is not guaranteed when data is heterogeneous [15]. Also, the divergence of clients' local objectives can lead the global model to converge to an erroneous point [26].

FedProx [15] regularizes the local training at the clients to restrict them to diverge from the parameters of the global model. This way the model takes small steps to learn but is better at handling heterogeneity in clients' local data distributions. However, a single global model usually struggle to have desirable performance for every client under significant client heterogeneity.

At the other end of the spectrum, instead of training and deploying the same global model, each client k can build a personalized model θ_k, which is the underlying motivation for pFL. Current methods in pFL can be categorized into two main groups: global model personalization and personalized model learning. In the prior group, a global model is first trained and then locally adapted to each client whereas the second category trains personalized models directly. Below we describe three methods that we explore in this paper covering both the categories.

Fine-tuning (FT) is a common method used for local adaptation, where a global model is trained using a federated mechanism till convergence. Then the model is retrained at every client using its local data for several epochs. It is simple to implement and compatible with any model architectures and FL strategies.

FedBN [16] aims to address covariate shifts between clients by leveraging the capabilities of batch normalization (BN) layers. It is achieved by maintaining local BN layer parameters and excluding them from federated aggregation.

FedAP [18] further considers client similarity and aggregates a customized model for every client, in additional to maintaining local BN layers. The similarity of clients' data distributions is measured by the distance between their BN layer information.

Both FedBN and FedAP have been applied in healthcare applications and demonstrate performance improvement in classification [16, 18]. We skip testing other advanced personalized FL methods from the literature, as this is orthogonal to our interest as stated in RQ2.

2.2 Uncertainty Quantification

After obtaining a local deep learning model θ_k for client k, the most basic way of measuring uncertainty is using softmax entropy, which is however known to be overconfident [8]. For neural networks, MC-Dropout and deep ensembles are two pervasive methods to introduce randomness and approximate the posterior parameter distribution.

Vanilla. In vanilla softmax entropy, only one inference is drawn and the uncertainty is measured by calculating the Shannon entropy of the predictive distribution,

$$\mathcal{H}(y|\mathbf{x}, \theta) = -\sum_{i=1}^{C} p_i \log(p_i), \tag{2}$$

where p_i represents the predicted probability of class i.

MC-Dropout (MCDrop). In MC-Dropout, dropout layers are added to the model and activated during inference, when T forward passes are run to obtain averaged output probabilities.

Deep Ensembles (Ensemble). Deep ensembles instead involve training M models with different weight initializations on the same data and the predictions from these models are averaged during inference.

Both MC-Dropout and deep ensemble sample an ensemble of predictions, $\mathcal{M} = \{P(y|\mathbf{x}, \theta_1), P(y|\mathbf{x}, \theta_2), \ldots, P(y|\mathbf{x}, \theta_{|\mathcal{M}|})\}$. Final predictions are obtained by their mean, and uncertainty can be measured by the predictive entropy

$$\mathcal{H}(y|x, D) := -\sum_{i=1}^{C} \left(\frac{1}{|\mathcal{M}|} \sum_{t=1}^{|\mathcal{M}|} a_t \right) \log \left(\frac{1}{|\mathcal{M}|} \sum_{t=1}^{|\mathcal{M}|} a_t \right), \tag{3}$$

where C is the number of classes, $|\mathcal{M}|$ is the size of the ensemble and $a_t = p(y = c_i|x, \theta_t)$ for each class c_i.

However, there is very limited work studying uncertainty estimation in FL. A recent study [17] demonstrates that prominent

approaches in centralized learning, including MC-Dropout and deep ensembles, can be extended to an FL setting with identical and independent distributed (IID) data distribution, without addressing data heterogeneity that is commonly present in healthcare applications.

2.3 Existing literature

Several existing works have been proposed to review and evaluate FL and pFL strategies. Chen et al. [2] presents a benchmark of several pFL strategies on 12 widely-used text and image datasets. However, their focus does not extend to realistic healthcare applications, which present unique challenges due to complex data heterogeneity and more importantly, the crucial need for uncertainty estimation. Terrail et al. [23] provides a benchmark of standard FL algorithms on healthcare datasets, but also did not touch on uncertainty quantification.

On the other hand, Xia et al. [27] evaluates several uncertainty quantification methods for capturing biosignal dataset shifts. But the behavior of uncertainty estimation for healthcare under the FL setting has not been studied and is the gap this paper addresses.

3 EXPERIMENTAL DESIGN

3.1 Datasets

We employ three realistic health datasets covering both cross-silo and cross-device scenarios [10]. These datasets are diverse in the number of samples and input modalities, with different levels of data heterogeneity, and are all tested on their natural partitions. Table. 1 provides an overview of the basic characteristics of the datasets.

Table 1: Summary of datasets and FL partition.

Dataset	Heart-Disease [1]	ISIC2019 [3, 4, 24]	PAMAP2 [20]
Modality	tabular features	image	time series
# Samples	740	23,247	2,869
# Clients	4	6	8
# Classes	2	8	8
Train Partition	199, 172, 30, 85	9930, 3163, 2691, 1807, 655, 351	173, 171, 179, 181, 176, 179, 188, 185
Test Partition	104, 89, 16, 45	2483, 791, 672, 452, 164, 88	174, 172, 179, 182, 177, 180, 188, 185
Input Dimension	13	$200 \times 200 \times 3$	1000×3

Heart-Disease. The Heart-Disease dataset [1] contains tabular information about patients collected in 4 hospitals located in three countries, with a binary classification task to predict the presence of heart disease. The baseline model is a 2-layer MLP with BN and Dropout layers.

ISIC2019. The ISIC2019 datasets [3, 4, 24] consist of dermoscopy images collected in 4 hospitals for melanoma class prediction. Since one hospital used 3 different imaging technologies throughout time, the data is partitioned into 6 clients in total. The task is a multi-class classification task among 8 different melanoma classes, and the baseline model is an EfficientNet-B0 [22] pretrained on ImageNet. Preprocessing steps follow the FLamby benchmark [23].

PAMAP2. The PAMAP2 dataset [20] contains data on different physical activities (such as walking, cycling, playing soccer, etc), measured by inertial measurement units (IMU), with a task to classify the activities. The baseline model is a 3-layer CNN model. We follow the preprocessing pipeline from Yuan et al. [28]. After preprocessing, the data contain 8 activity classes performed by 8 subjects, each acting as a client.

In these real-world health datasets, clients exhibit universal data heterogeneity, including an imbalanced number of samples, distinct label distributions, different patient demographics and technologies used for data collection. A more detailed illustration of data label distribution can be found in Appendix. A.

3.2 Settings

We evaluate the quality of the uncertainty estimated by three uncertainty methods, namely vanilla softmax entropy, MC-Dropout and deep ensembles in FL. For federated MC-Dropout, we activate dropout layers in the models of all clients during inference. As for federated deep ensembles, we train an ensemble of global models, which has shown superior performance compared to other variations, such as ensemble of local models, as demonstrated in a previous study in IID FL [17]. They are applied in four FL strategies with increasing personalization degree, FedAvg, FedProx, FedBN and FedAP, as well as combined with FT.

We adopted two settings to measure the model's ability to capture potential misdiagnoses [5]. In the first scenario, we measure the model's ability to distinguish correctly and incorrectly predicted samples, i.e. ***misclassification detection***. We calculate the area under the receiver operating curve (AUROC) for classification based on the predictive entropy. In the second scenario, we discard the most uncertain samples from the test dataset and solely evaluate the prediction performance of the remaining data, i.e., ***selective prediction*** or abstention. This scenario allows the model to refrain from making unconfident predictions and leave them for human inspection. Observing a consistent comparative performance regardless of the chosen thresholds, we choose to discard 40% samples in the results for a clear distinction.

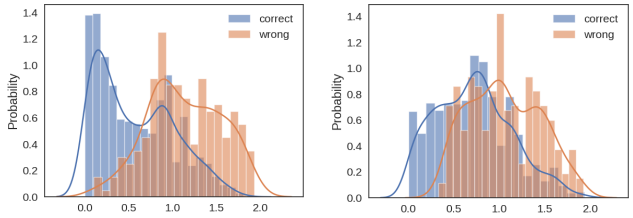
We train on one NVIDIA A100 GPU, and adopt $T = 1000$ for MC-Dropout and $T = 5$ for deep ensembles, due to empirical studies revealing these parameters to have adequate and comparable performance in centralised settings. Our detailed experimental setup is concluded in Appendix B.

4 RESULTS AND FINDINGS

4.1 Uncertainty Quantification with Data Heterogeneity

To answer **RQ1**, we first evaluate the uncertainty methods under the most common FL strategy FedAvg, to determine their effectiveness within the context of FL when faced with statistical data heterogeneity in health applications.

Results indicate that MC-dropout and deep ensembles can be extended to fit real-world health applications in federated learning. Nevertheless, the uncertainty estimation obtained by federated models degrades along with the accuracy when compared with centralized models. The difference in uncertainty quality can be observed in Fig. 1. The reported values of the two settings can be found in Appendix C, along with classification accuracy.



(a) Centralized (b) FedAvg

Figure 1: Uncertainty distribution of FedAvg model compared to the centralized model on PAMAP2 dataset.

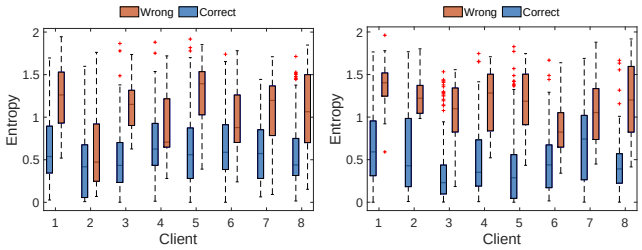
Table 2: Uncertainty quality of FL and pFL models on. The best-performing FL strategy is highlighted in bold, the centralized performance is in violet for reference, and the top-performing uncertainty method for each strategy is underlined.

Dataset	Strategy	Misclassification Detection			Selective Prediction		
		Vanilla	MCDrop	Ensemble	Vanilla	MCDrop	Ensemble
Heart-Disease	Centralized	0.621	0.620	0.618	0.795	0.796	0.810
	FedAvg	0.596	0.600	0.580	0.686	0.686	0.684
	FedProx	0.577	0.584	0.575	0.682	0.682	0.683
	FedBN	0.685	0.690	0.692	0.859	0.857	0.863
	FedAP	**0.687**	**0.692**	**0.699**	**0.863**	**0.863**	0.867
	FedAvg-FT	0.682	0.681	0.697	0.851	0.854	0.865
	FedProx-FT	0.665	0.664	0.683	0.848	0.851	**0.877**
ISIC2019	Centralized	0.748	0.748	0.766	0.860	0.860	0.917
	FedAvg	0.804	0.804	0.804	0.827	0.827	0.847
	FedProx	0.831	0.830	0.835	0.895	0.895	0.916
	FedBN	0.817	0.817	0.822	0.898	0.905	0.951
	FedAP	0.832	**0.832**	0.847	**0.933**	**0.933**	0.959
	FedAvg-FT	0.839	0.831	0.866	0.920	0.867	0.953
	FedProx-FT	**0.841**	0.830	**0.869**	0.908	0.891	**0.962**
PAMAP2	Centralized	0.791	0.821	0.816	0.908	0.922	0.923
	FedAvg	0.769	0.817	0.788	0.866	0.894	0.877
	FedProx	0.771	0.812	0.775	0.868	0.890	0.866
	FedBN	0.762	0.803	0.794	0.873	0.896	0.900
	FedAP	0.851	0.874	0.866	0.946	0.965	0.964
	FedAvg-FT	**0.878**	**0.895**	**0.907**	**0.980**	**0.986**	0.991
	FedProx-FT	0.860	0.888	0.902	0.971	0.981	0.991
	FedBN-FT	0.870	0.890	0.902	0.973	0.981	**0.994**
	FedAP-FT	0.874	0.885	**0.907**	0.977	0.980	0.992

4.2 Uncertainty Quantification with Personalization

Personalization has been investigated in FL to improve local performance under data heterogeneity. However, it remains unclear if these methods can also enhance uncertainty estimation. Thus, **RQ2** and our second set of experiments aim to verify this on the selected healthcare datasets. We evaluate various personalization methods, including FL, pFL strategies, and fine-tuning.

Table. 2 presents the evaluated quality of estimated uncertainty under the two settings. Each reported value is the mean of 5 runs, and the standard deviations (with details in Appendix C) are very small, mostly in the order of 10^{-2} or below. We can see great improvements brought by personalization of up to 12% in misclassification detection and 19% in selective prediction. The best estimated uncertainty is always produced by personalized methods. Surprisingly, personalized methods can even outperform centralized models almost consistently in all datasets. Fig. 2 illustrates in detail how the quality of uncertainty improves after personalization.



(a) FedAvg (b) FedAvg-FT

Figure 2: Misclassification detection results of FedAvg model on PAMAP2 dataset, with and without fine-tuning.

Both pFL strategies and post-training local adaptation prove effective in tackling data heterogeneity, with FedAvg-FT, FedProx-FT and FedAP being the most promising, while their comparative superiority depends partially on the specific application. Notably, FedAP consistently outperforms FedBN, highlighting the effectiveness of client clustering in mitigating data heterogeneity. On the other hand, FedBN fails to improve over FedAvg in PAMAP2 dataset, where feature shift is less severe, likely due to its inability to handle label distribution heterogeneity. In PAMAP2, fine-tuning also improves upon the personalized methods, suggesting the need to customize appropriate personalization approach based on the specific circumstances.

4.3 Comparative Analysis of Uncertainty Methods

To answer **RQ3**, we now investigate the effectiveness of the chosen uncertainty quantification methods in our realistic FL settings. From the tables, we can observe that federated deep ensembles have the best performance across datasets and personalization strategies.

In Fig. 3, it is evident that deep ensembles perform better at every percentage of samples removed in selective prediction. Not only does it exhibit consistently higher accuracy, but it also demonstrates superior uncertainty estimation quality. This is exemplified by a notable 5% accuracy improvement when discarding only 5% of the samples and a further 6% increase at around 20% discard rate.

4.4 Summary of Findings

After analyzing the results from these experiments, we summarize our findings regarding the research questions. Firstly, uncertainty methods can be extended into non-IID FL for preventing misdiagnoses in healthcare applications. However, the quality of uncertainty degrades along with classification accuracy. On the other

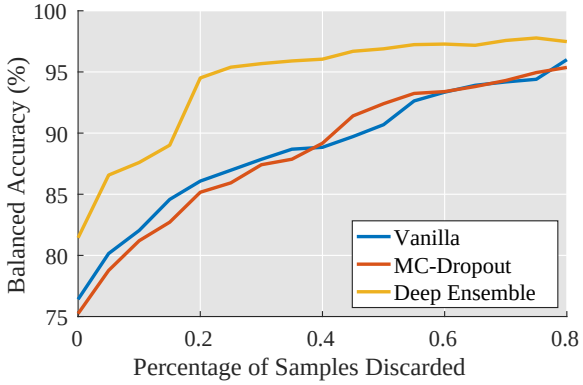


Figure 3: Evolution of local performance of FedProx-FT on ISIC2019 dataset as the most uncertain samples are removed.

hand, personalization proves effective in enhancing uncertainty estimation, with up to 19% improvement. Among the uncertainty methods, federated deep ensembles perform consistently well in this setting.

5 DISCUSSION AND FUTURE WORK

This paper presents the first evaluation of different uncertainty quantification methods in FL settings under data heterogeneity using real-world health datasets. Experimental results show that personalization in FL not only improves classification accuracy but also increases the quality of estimated uncertainty. Thus personalization is a promising research direction in local client deployment and uncertainty quantification for healthcare applications. We found that federated deep ensembles perform consistently better than the other uncertainty methods. However, it is notable that this advantage is at the expense of introducing higher computational costs. In our future work, we will further investigate more cost-efficient uncertainty quantification methods and make them more feasible for federated learning.

REFERENCES

[1] Janosi Andras, Steinbrunn William, Pfisterer Matthias, and Detrano Robert. 1988. Heart disease data set. (1988).

[2] Daoyuan Chen, Dawei Gao, Weirui Kuang, Yaliang Li, and Bolin Ding. [n. d.]. pFL-Bench: A Comprehensive Benchmark for Personalized Federated Learning. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[3] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. 2018. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). In *IEEE International Symposium on Biomedical Imaging*. 168–172.

[4] Marc Combalia, Noel CF Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Allan C Halpern, Susana Puig, et al. 2019. Bcn20000: Dermoscopic lesions in the wild. *arXiv preprint arXiv:1908.02288* (2019).

[5] Marc Combalia, Ferran Hueto, Susana Puig, Josep Malvehy, and Veronica Vilaplana. 2020. Uncertainty estimation in deep neural networks for dermoscopic image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 744–745.

[6] Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, et al. 2018. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine* 24, 9 (2018), 1342–1350.

[7] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*. PMLR, 1050–1059.

[8] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*. PMLR, 1321–1330.

[9] Jing Han, Tong Xia, Dimitris Spathis, Erika Bondareva, Chloë Brown, Jagmohan Chauhan, Ting Dang, Andreas Grammenos, Apinan Hasthanasombat, Andres Floto, et al. 2022. Sounds of COVID-19: exploring realistic performance of audio-based digital testing. *NPJ digital medicine* 5, 1 (2022), 1–9.

[10] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. 2021. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning* 14, 1–2 (2021), 1–210.

[11] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*.

[12] Benjamin Kompa, Jasper Snoek, and Andrew L Beam. 2021. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digital Medicine* 4, 1 (2021), 1–6.

[13] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems* 30 (2017).

[14] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. 2020. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine* 37, 3 (2020), 50–60.

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

[16] Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. 2021. FedBN: Federated Learning on Non-IID Features via Local Batch Normalization. In *International Conference on Learning Representations*.

[17] Florian Linsner, Linara Adilova, Sina Däubener, Michael Kamp, and Asja Fischer. 2021. Approaches to Uncertainty Quantification in Federated Deep Learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 128–145.

[18] Wang Lu, Jindong Wang, Yiqiang Chen, Xin Qin, Renjun Xu, Dimitrios Dimitriadis, and Tao Qin. 2022. Personalized federated learning with adaptive batchnorm for healthcare. *IEEE Transactions on Big Data* (2022).

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

[20] Attila Reiss and Didier Stricker. 2012. Introducing a new benchmarked dataset for activity monitoring. In *2012 16th international symposium on wearable computers*. IEEE, 108–109.

[21] Alysa Ziying Tan, Han Yu, Li zhen Cui, and Qiang Yang. 2021. Towards Personalized Federated Learning. *IEEE Transactions on Neural Networks and Learning Systems* (2021).

[22] Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*. 6105–6114.

[23] Jean Ogier du Terrail, Samy-Safwan Ayed, Edwige Cyffers, Felix Grimberg, Chaoyang He, Regis Loeb, Paul Mangold, Tanguy Marchand, Othmane Marfoq, Erum Mushtaq, et al. 2022. FLamby: Datasets and Benchmarks for Cross-Silo Federated Learning in Realistic Healthcare Settings. *arXiv preprint arXiv:2210.04620* (2022).

[24] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data* (2018), 1–9.

[25] José R Villar, Silvia González, Javier Sedano, Camelia Chira, and Jose M Trejo-Gabriel-Galan. 2015. Improving human activity recognition and its application in early stroke diagnosis. *International journal of neural systems* 25, 04 (2015), 1450036.

[26] Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. 2020. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems* 33 (2020), 7611–7623.

[27] Tong Xia, Jing Han, and Cecilia Mascolo. 2022. Benchmarking uncertainty quantification on biosignal classification tasks under dataset shift. In *Multimodal AI in healthcare: A paradigm shift in health intelligence*. Springer, 347–359.

[28] Hang Yuan, Shing Chan, Andrew P Creagh, Catherine Tong, David A Clifton, and Aiden Doherty. 2022. Self-supervised learning for human activity recognition using 700,000 person-days of wearable data. *arXiv preprint arXiv:2206.02909* (2022).

A DATASET DESCRIPTION

The label distributions of the three datasets are concluded in Fig. 4, Fig. 5 and Fig. 6.

B IMPLEMENTATION DETAILS

Table. 3 presents the chosen setup for training and evaluation. The optimizer used in all experiments is Adam [11]. The chosen μ for FedProx is set to be 0.01. The number of federated training rounds is calculated using the same method as in the implementation of the FLamby benchmark [23], and is same for all the strategies. The number of rounds T is calculated as

$$T = n_{epoches}^{pooled} \cdot \lfloor N/K/B/E \rfloor, \tag{4}$$

where $n_{epoches}^{pooled}$ is the number of epoches required to train the centralized model, N is the total number of training samples, K is the number of clients, B is the batch size and E is the number of local epoches.

The implementation of the 4 FL and pFL strategies follows an open-source federated learning codebase based on PyTorch developed by Microsoft[1]. The code is re-writed to be compatible with the chosen dataset, models and tasks, and also modified to be more flexible for extensive experiments. The uncertainty methods are implemented by this work orthogonal to the strategies and allows future extensions.

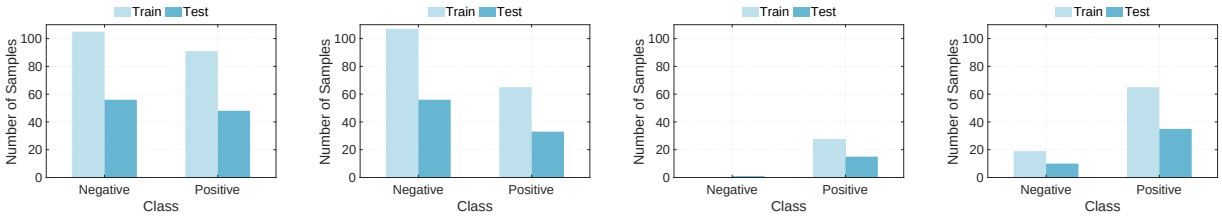
For each reported value in the table, we run 5 experiments to get the mean and standard deviation. Whereas for federated deep ensembles, we randomly sample 5 models from 10 trained models (4 out of 5 for ISIC2019) due to computational constraints.

C SUPPLEMENTARY RESULTS

Table. 4 and Table. 5 presents the performance of classification and uncertainty quantification comparing FedAvg and centralized learning.

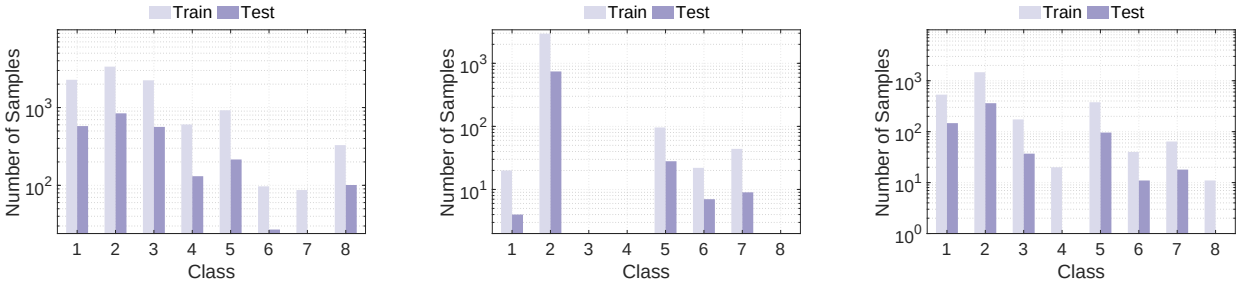
Table. 6 and Table. 7 presents the performance of classification and uncertainty quantification comparing standard FL and pFL.

[1]https://github.com/microsoft/PersonalizedFL

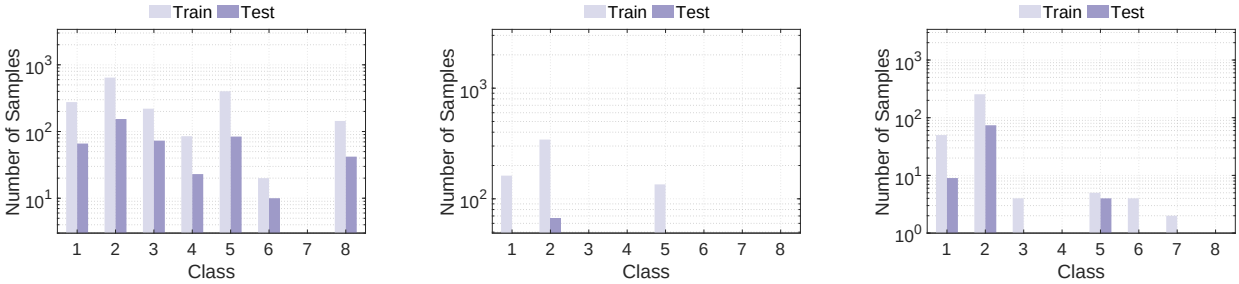


(a) Client 1 (b) Client 2 (c) Client 3 (d) Client 4

Figure 4: Label distribution of the 4 clients in Heart-Disease dataset.



(a) Client 1 (b) Client 2 (c) Client 3

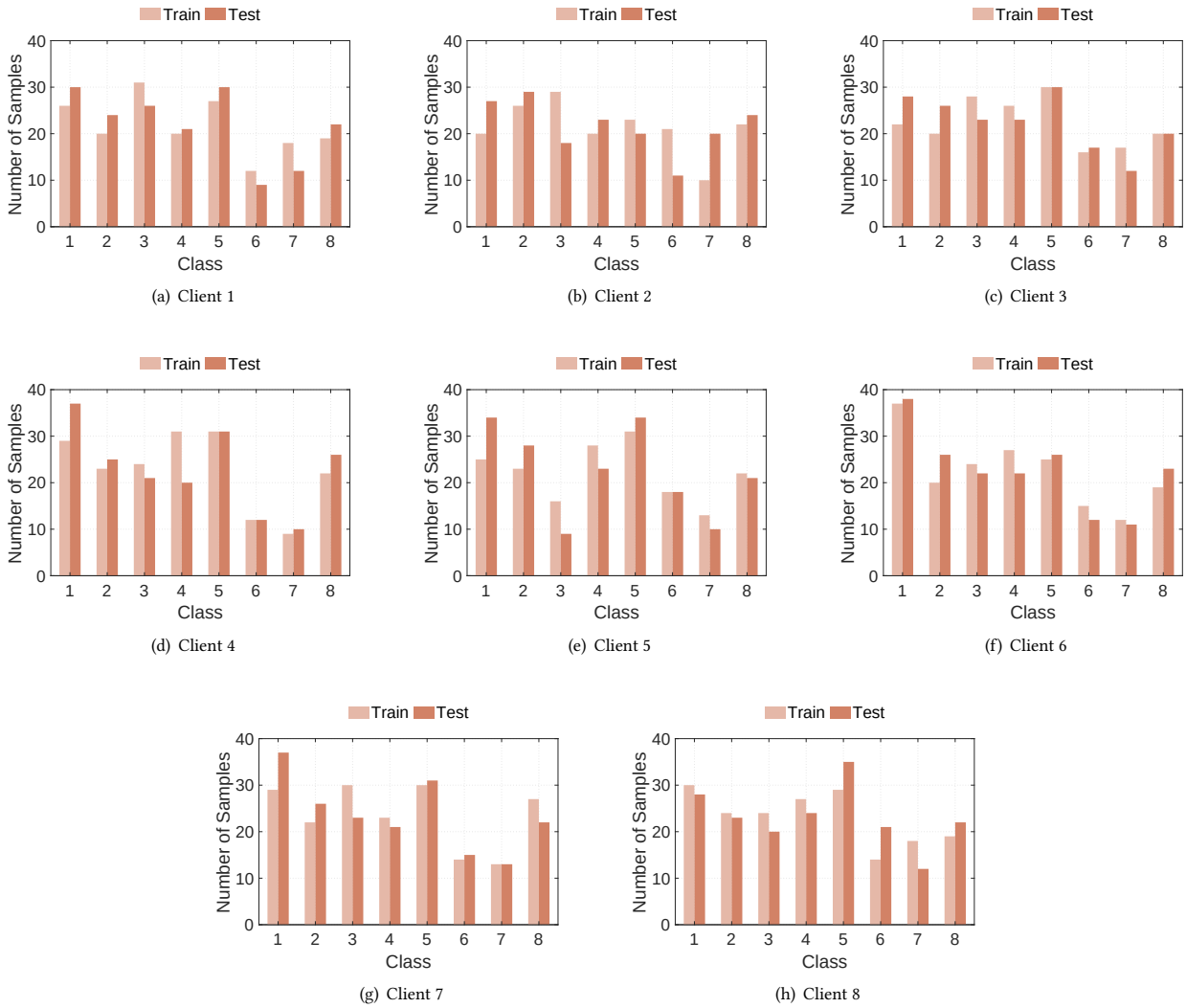


(d) Client 4 (e) Client 5 (f) Client 6

Figure 5: Label distribution of the 6 clients in ISIC-2019 dataset in log-scale.

Table 3: Implementation details.

Dataset	Heart-Disease	ISIC2019	PAMAP2	PhysioNet-2016
Model	2-layer MLP	EfficientNet-B0	CNN	CNN
Batch Size	4	64	32	64
Learning Rate	0.001	0.005	0.0003	0.0005
# Local Iters	50	20	20	30
# Rounds	30	47	10	22
Metric	Accuracy	Balanced Accuracy	Macro F1-score	Accuracy
Loss	BCE Loss	Weighted Focal Loss	CE Loss	BCE Loss



(a) Client 1 (b) Client 2 (c) Client 3

(d) Client 4 (e) Client 5 (f) Client 6

(g) Client 7 (h) Client 8

Figure 6: Label distribution of the 8 clients in PAMAP2 dataset.

Table 4: Performance of centralized and federated models on the pooled test set.

Dataset	Strategy	Accuracy		
		Vanilla	MCDropout	Deep Ensemble
Heart-Disease	Centralized	**0.806 ± 0.009**	**0.806 ± 0.008**	**0.756 ± 0.016**
	FedAvg	0.680 ± 0.027	0.680 ± 0.027	0.644 ± 0.009
ISIC-2019	Centralized	**0.753 ± 0.016**	**0.754 ± 0.015**	**0.823 ± 0.005**
	FedAvg	0.704 ± 0.021	0.704 ± 0.021	0.728 ± 0.005
PAMAP2	Centralized	**0.852 ± 0.009**	**0.849 ± 0.009**	**0.880 ± 0.005**
	FedAvg	0.749 ± 0.023	0.760 ± 0.024	0.770 ± 0.005

Table 5: Uncertainty quality of centralized and federated models on the pooled test set.

Dataset	Strategy	Misclassification detection			Selective prediction		
		Vanilla	MC-Dropout	Deep Ensembles	Vanilla	MC-Dropout	Deep Ensembles
Heart-Disease	Centralized	**0.672 ± 0.034**	**0.669 ± 0.032**	**0.666 ± 0.016**	**0.872 ± 0.012**	**0.872 ± 0.012**	**0.817 ± 0.004**
	FedAvg	0.613 ± 0.024	0.613 ± 0.027	0.609 ± 0.010	0.753 ± 0.009	0.751 ± 0.011	0.685 ± 0.005
ISIC-2019	Centralized	0.762 ± 0.015	0.761 ± 0.015	0.776 ± 0.002	**0.871 ± 0.013**	**0.873 ± 0.014**	**0.926 ± 0.003**
	FedAvg	**0.810 ± 0.006**	**0.810 ± 0.006**	**0.807 ± 0.003**	0.831 ± 0.042	0.830 ± 0.043	0.863 ± 0.015
PAMAP2	Centralized	**0.831 ± 0.015**	**0.861 ± 0.010**	**0.881 ± 0.003**	**0.959 ± 0.007**	**0.971 ± 0.003**	**0.984 ± 0.002**
	FedAvg	0.740 ± 0.029	0.763 ± 0.025	0.763 ± 0.011	0.843 ± 0.029	0.878 ± 0.020	0.877 ± 0.007

Table 6: Accuracy of FL and pFL models, with mean and standard deviation across 5 runs. The best-performing FL strategy is highlighted in bold, the centralized performance is in violet for reference, and the top-performing uncertainty method for each strategy is underlined.

Dataset	Strategy	Accuracy		
		Vanilla	MC-Dropout	Deep Ensembles
Heart-Disease	Centralized	0.736 ± 0.014	0.737 ± 0.014	0.743 ± 0.008
	FedAvg	0.648 ± 0.026	0.648 ± 0.028	0.665 ± 0.008
	FedProx	0.665 ± 0.022	0.664 ± 0.023	0.670 ± 0.009
	FedBN	0.771 ± 0.007	0.770 ± 0.006	0.779 ± 0.006
	FedAP	0.775 ± 0.007	0.775 ± 0.007	0.774 ± 0.005
	FedAvg-FT	0.772 ± 0.009	0.773 ± 0.010	0.781 ± 0.009
	FedProx-FT	**0.779 ± 0.008**	**0.780 ± 0.009**	**0.791 ± 0.005**
ISIC-2019	Centralized	0.736 ± 0.007	0.737 ± 0.006	0.797 ± 0.004
	FedAvg	0.710 ± 0.018	0.710 ± 0.018	0.725 ± 0.004
	FedProx	0.754 ± 0.015	0.754 ± 0.015	0.771 ± 0.003
	FedBN	0.756 ± 0.047	0.756 ± 0.047	0.808 ± 0.005
	FedAP	**0.794 ± 0.017**	**0.794 ± 0.017**	**0.817 ± 0.005**
	FedAvg-FT	0.760 ± 0.033	0.756 ± 0.020	0.783 ± 0.005
	FedProx-FT	0.757 ± 0.024	0.750 ± 0.020	0.796 ± 0.013
PAMAP2	Centralized	0.802 ± 0.023	0.796 ± 0.027	0.832 ± 0.002
	FedAvg	0.780 ± 0.014	0.782 ± 0.017	0.787 ± 0.007
	FedProx	0.775 ± 0.006	0.778 ± 0.010	0.782 ± 0.003
	FedBN	0.783 ± 0.013	0.781 ± 0.014	0.808 ± 0.006
	FedAP	0.835 ± 0.009	0.842 ± 0.008	0.864 ± 0.003
	FedAvg-FT	**0.877 ± 0.009**	**0.870 ± 0.012**	**0.894 ± 0.003**
	FedProx-FT	0.869 ± 0.003	0.866 ± 0.004	0.892 ± 0.003
	FedBN-FT	0.870 ± 0.008	0.867 ± 0.008	0.897 ± 0.003
	FedAP-FT	0.873 ± 0.006	0.868 ± 0.005	0.897 ± 0.003

Table 7: Uncertainty quality of FL and pFL models, with mean and standard deviation across 5 runs. The best-performing FL strategy is highlighted in bold, the centralized performance is in violet for reference, and the top-performing uncertainty method for each strategy is underlined.

Dataset	Strategy	Misclassification detection			Selective prediction		
		Vanilla	MC-Dropout	Deep Ensembles	Vanilla	MC-Dropout	Deep Ensembles
Heart-Disease	Centralized	0.621 ± 0.038	0.620 ± 0.035	0.618 ± 0.009	0.795 ± 0.021	0.796 ± 0.021	0.810 ± 0.005
	FedAvg	0.596 ± 0.027	0.600 ± 0.030	0.580 ± 0.011	0.686 ± 0.005	0.686 ± 0.007	0.684 ± 0.003
	FedProx	0.577 ± 0.023	0.584 ± 0.024	0.575 ± 0.011	0.682 ± 0.008	0.682 ± 0.008	0.683 ± 0.000
	FedBN	0.685 ± 0.025	0.690 ± 0.026	0.692 ± 0.014	0.859 ± 0.014	0.857 ± 0.016	0.863 ± 0.005
	FedAP	**0.687 ± 0.024**	**0.692 ± 0.018**	**0.699 ± 0.013**	**0.863 ± 0.009**	**0.863 ± 0.009**	0.867 ± 0.005
	FedAvg-FT	0.682 ± 0.019	0.681 ± 0.017	0.697 ± 0.012	0.851 ± 0.003	0.854 ± 0.005	0.865 ± 0.003
	FedProx-FT	0.665 ± 0.010	0.664 ± 0.015	0.683 ± 0.013	0.848 ± 0.004	0.851 ± 0.010	**0.877 ± 0.007**
ISIC-2019	Centralized	0.748 ± 0.016	0.748 ± 0.016	0.766 ± 0.003	0.860 ± 0.013	0.860 ± 0.013	0.917 ± 0.003
	FedAvg	0.804 ± 0.007	0.804 ± 0.007	0.804 ± 0.004	0.827 ± 0.028	0.827 ± 0.029	0.847 ± 0.010
	FedProx	0.831 ± 0.005	0.830 ± 0.005	0.835 ± 0.001	0.895 ± 0.013	0.895 ± 0.013	0.916 ± 0.004
	FedBN	0.817 ± 0.026	0.817 ± 0.026	0.822 ± 0.009	0.898 ± 0.060	0.905 ± 0.061	0.951 ± 0.003
	FedAP	0.832 ± 0.027	**0.832 ± 0.027**	0.847 ± 0.007	**0.933 ± 0.026**	**0.933 ± 0.026**	0.959 ± 0.004
	FedAvg-FT	0.839 ± 0.010	0.831 ± 0.008	0.866 ± 0.005	0.920 ± 0.021	0.867 ± 0.023	0.953 ± 0.017
	FedProx-FT	**0.841 ± 0.013**	0.830 ± 0.021	**0.869 ± 0.007**	0.908 ± 0.046	0.891 ± 0.029	**0.962 ± 0.002**
PAMAP2	Centralized	0.791 ± 0.016	0.821 ± 0.020	0.816 ± 0.009	0.908 ± 0.021	0.922 ± 0.020	0.923 ± 0.003
	FedAvg	0.769 ± 0.020	0.817 ± 0.011	0.788 ± 0.006	0.866 ± 0.016	0.894 ± 0.017	0.877 ± 0.008
	FedProx	0.771 ± 0.009	0.812 ± 0.018	0.775 ± 0.006	0.868 ± 0.016	0.890 ± 0.010	0.866 ± 0.006
	FedBN	0.762 ± 0.037	0.803 ± 0.016	0.794 ± 0.012	0.873 ± 0.016	0.896 ± 0.014	0.900 ± 0.002
	FedAP	0.851 ± 0.015	0.874 ± 0.013	0.866 ± 0.007	0.946 ± 0.009	0.965 ± 0.007	0.964 ± 0.003
	FedAvg-FT	**0.878 ± 0.012**	**0.895 ± 0.009**	**0.907 ± 0.004**	**0.980 ± 0.007**	**0.986 ± 0.006**	0.991 ± 0.002
	FedProx-FT	0.860 ± 0.019	0.888 ± 0.010	0.902 ± 0.003	0.971 ± 0.004	0.981 ± 0.003	0.991 ± 0.002
	FedBN-FT	0.870 ± 0.018	0.890 ± 0.009	0.902 ± 0.003	0.973 ± 0.010	0.981 ± 0.007	**0.994 ± 0.001**
	FedAP-FT	0.874 ± 0.016	0.885 ± 0.015	0.907 ± 0.007	0.977 ± 0.006	0.980 ± 0.006	0.992 ± 0.001