FedEE: Uncertainty-Aware Personalized Federated Learning for Realistic Healthcare Applications

Yuwei Zhang[1] YZ798@CAM.AC.UK
Tong Xia[1*] TX229@CAM.AC.UK
Abhirup Ghosh[2] A.GHOSH.1@BHAM.AC.UK
Cecilia Mascolo[1] CM542@CAM.AC.UK
[1] *University of Cambridge, UK* [2] *University of Birmingham, UK*
[*] *Corresponding author*

Abstract

Healthcare applications require accurate and uncertainty-aware machine learning models, providing confidence rather than only black-box predictions. However, training such deep learning models with insufficient data at individual sites (e.g., hospitals) poses a challenge. Federated learning (FL) mitigates this by allowing data holders to collaboratively train models without sharing sensitive health data. Yet, we identify two major realistic challenges when building uncertainty estimates in FL, *severe data heterogeneity* and *high computational overhead*. This paper proposes **FedEE**, an uncertainty-aware and efficient personalized FL framework for realistic healthcare applications. FedEE achieves an efficient way of ensembling by incorporating lightweight early exit blocks into a single backbone model. These blocks are personalized for each client to tackle data heterogeneity. Experiments with four FL strategies and three datasets demonstrate that FedEE achieves up to a 15% improvement in uncertainty estimation from vanilla softmax entropy and is competitive with expensive baselines, showcasing in the order of 5× improved efficiency with a 5-member ensemble.

Keywords: Uncertainty quantification, federated learning, healthcare, data heterogeneity

Data and Code Availability We use the following open data: (1) the PAMAP2 dataset (Reiss and Stricker, 2012) (2) The ISIC2019 datasets (Tschandl et al., 2018; Codella et al., 2018; Combalia et al., 2019) (3) the PhysioNet-2016 dataset (Liu et al., 2016; Goldberger et al., 2000). The code can be found in our Github Repository.

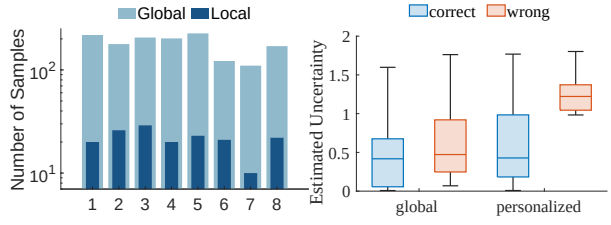
Institutional Review Board (IRB) This study obtained IRB approval by the Ethics Committee of the Department of Computer Science and Technology at the University of Cambridge to work with the public data described.

1. Introduction

In line with the impact on other domains of science, deep learning has become popular in medical diagnosis and health monitoring. However, clinicians often hesitate to trust results from black-box deep learning models mainly due to the potential of unaware incorrect predictions. One way to tackle this problem is by incorporating a calibrated confidence score along with the model prediction (Bhatt et al., 2021). This uncertainty awareness is crucial but challenging, especially when a model is trained with limited data (leading to overfitting), which is the case, for example, if a hospital trains solely using its own data.

Modern deep learning architectures rely on large amounts of data to learn robust and generalizable models (Alzubaidi et al., 2021). The standard way of aggregating data across multiple data sources is difficult for sensitive health data given the privacy restrictions. Federated Learning (FL) (Kairouz et al., 2021) emerges as a promising solution, enabling the data holders (called clients) to collaboratively train a model while keeping their private data local.

The above arguments suggest that a practical deep learning system for health needs to provide uncertainty estimates for predictions and be developed within an FL setting. While uncertainty estimation has been extensively studied in centralized machine learning (Abdar et al., 2021), systems with both capabilities are rare in general and absent in the con-



(a) data distribution (b) uncertainty estimate

Figure 1: Uncertainty estimation for an FL client with a distinct data distribution. When local data distribution differs from the global distribution (a), the uncertainty quantified by the global model is less useful than that of the personalized model in distinguishing correct/incorrect predictions (b).

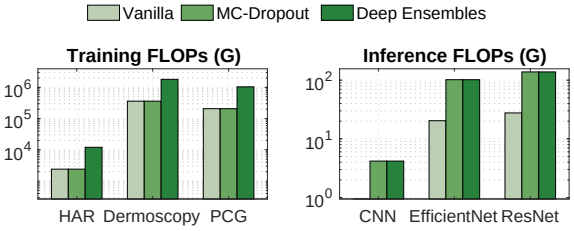


Figure 2: Computation cost of common methods on three healthcare applications, using one round FL training and one batch inference.

text of health applications. Our paper fills this research gap by uncovering and addressing the non-trivial challenges hindering uncertainty quantification in decentralized settings.

Challenges. *Firstly*, health data across clients is highly heterogeneous due to factors like geographic diversity of disease prevalence and variations in data collection technologies. Consequently, a single model often fails to produce accurate predictions across all clients (Tan et al., 2021). We find that uncertainty estimates also suffer from such heterogeneity and simple adaptation of uncertainty quantification methods with FL aggregation yields poor performance (Zhang et al., 2023). Figure 1 illustrates this issue, where a global model struggles with inadequate uncertainty estimation for data that deviates from the global distribution, while a personalized model gives high-quality uncertainty that could differentiate correct and incorrect local client predictions.

Secondly, popular uncertainty estimation methods in deep learning either learn multiple versions of a model, as in ensembles (Lakshminarayanan et al., 2017) and Bayesian networks (MacKay, 1992), or introduce stochastic processes at inference as in Monte-Carlo dropout (Gal and Ghahramani, 2016). All these methods are compute and memory intensive. This is evident in Figure 2, which shows the overhead for applications like human activity recognition (HAR), dermoscopy melanoma prediction (dermoscopy), and PCG heart disease classification (PCG), with dataset details in Section 5.1. In FL, this overhead is amplified due to longer convergence time and the need for network communication (Qiu et al., 2023), raising entry barriers, especially for smaller clinics or developing regions with limited computational resources.

This work. To address the above challenges we propose **FedEE**, an uncertainty-aware and efficient personalized FL framework for realistic healthcare applications. FedEE employs Early Exit Ensembles (EEE), incorporating lightweight early exit blocks with minimal parameters to form an ensemble of weight-sharing sub-networks from the backbone model. To handle health data heterogeneity, FedEE takes a personalized approach by training the early exit blocks only using local data, while aggregating the backbone deep learning model across clients. The computational overhead is substantially diminished with the efficient ensembling, as FedEE requires only one backbone model during training and one forward pass during inference.

We evaluate FedEE against three baselines using four FL strategies on three real-world multi-site healthcare datasets with natural partitions. These datasets encompass a range of classification applications, including heart disease detection, human activity recognition, and melanoma class classification, involving various types of input data, such as audio, images, and time series. To evaluate the effectiveness of the uncertainty qualification methods, we design two evaluation tasks: misclassification detection and selective prediction. Additionally, we measure the memory, computation and communication costs of the methods. The results highlight FedEE's comparable or superior performance to baselines with significantly improved efficiency and sustainability.

Our main contributions are summarized as follows:

- For the first time, we study the problem of uncertainty quantification using decentralized health data. We identify that data heterogeneity and high computation cost are two major challenges hindering trustworthy deep learning for health.

- We propose a efficient and uncertainty-aware personalized FL framework, FedEE, leveraging early exit ensembles to address the 2 above challenges.
- Extensive experiments demonstrate that our method enhances uncertainty estimation in FL with data heterogeneity by up to 15%, while inducing minimal overhead (approximately 5× higher efficiency than baselines).

2. Related Work

2.1. Federated Learning

Despite advancements in deep learning research for healthcare, one of the crucial challenges remaining is data availability (Kelly et al., 2019). FL offers a promising solution by enabling collaborative training without transferring sensitive data. However, data heterogeneity, prevalent in healthcare, has been a major concern prohibiting learning a single strong global model for all clients due to poor convergence and objective inconsistency (Li et al., 2020; Wang et al., 2020b). Efforts have focused on two directions: i) enhancing the FL process for global model robustness and generalizability, and ii) personalizing local models based on individual data.

The first set of methods includes data-driven approaches (Zhao et al., 2018; Xia et al., 2024b), federated optimization methods (Reddi et al., 2020), client selection and server aggregation methods (Wang et al., 2020a; Xia et al., 2023), and regularization methods preventing model divergence (Li et al., 2020). The second set involves personalizing the global model via transfer learning (Kairouz et al., 2021; Mansour et al., 2020), or training personalized models locally with benefits from other clients, for example through parameter decoupling (Li et al., 2021b), or client similarities (Lu et al., 2022). While some personalized methods like FedBN (Li et al., 2021b) and FedAP (Lu et al., 2022) show great promise with non-IID health data, existing work often overlooks the necessity for providing high quality uncertainty estimation in addition to high predictive performance.

2.2. Uncertainty Quantification

A trustworthy health model needs to convey uncertainty when predictions may be inaccurate, prompting human intervention and improving risk management. However, deep learning models are typically trained and evaluated for classification metrics like accuracy, AUC, sensitivity, and specificity, which may not truly reflect clinical applicability (Kelly et al., 2019). Communicating model uncertainty is crucial for trustworthy healthcare applications to reduce overconfident misdiagnoses but often overlooked.

Deep learning models traditionally express confidence through the entropy value of predicted probabilities, but this approach is known to be overconfident (Guo et al., 2017). Bayesian Neural Networks (MacKay, 1992) treat weights and outputs as random variables instead of single values, but are computationally expensive and intractable for neural networks, and data-intensive in training (Hernández-Lobato and Adams, 2015; Xia et al., 2024a). For neural networks and deep learning, researchers has proposed Frequentist methods to introduce randomness and approximate the posterior parameter distribution, such as Monte Carlo (MC) Dropout (Gal and Ghahramani, 2016) and deep ensembles (Lakshminarayanan et al., 2017). Their direct adoption in FL have been explored under IID settings (Linsner et al., 2021) and show promising potential, but can be unrealistic in practice due to high memory, computation, and communication demands. They also fail to tackle the data heterogeneity commonly present in healthcare applications and affecting uncertainty estimation.

Some approaches have explored practical ensembling and bayesian methods in federated learning for accuracy improvement, but they did not focus on estimated uncertainty and many had additional requirements unsuitable for healthcare applications (Chen and Chao, 2021; Shi et al., 2021; Zhu et al., 2023), including extra server datasets and a large number of clients for sampling the ensemble.

Early exit ensembles (EEE) (Qendro et al., 2021) have proven effective and efficient in uncertainty estimation for centralized learning. Initially introduced to prevent overfitting by dynamically altering the computational graph of a neural network, early exits have been extended to quantify uncertainty in medical imaging and biosignal classification (Qendro et al., 2021; Campbell et al., 2022). We build upon this early exits approach and for the first time apply it in the context of federated learning for efficient and personalized uncertainty estimation.

3. Problem Formulation

In this section, we introduce the preliminaries and annotations to be used throughout the paper, followed by formulating our problem and goals. Since a model might consist of multiple members of a neural net-

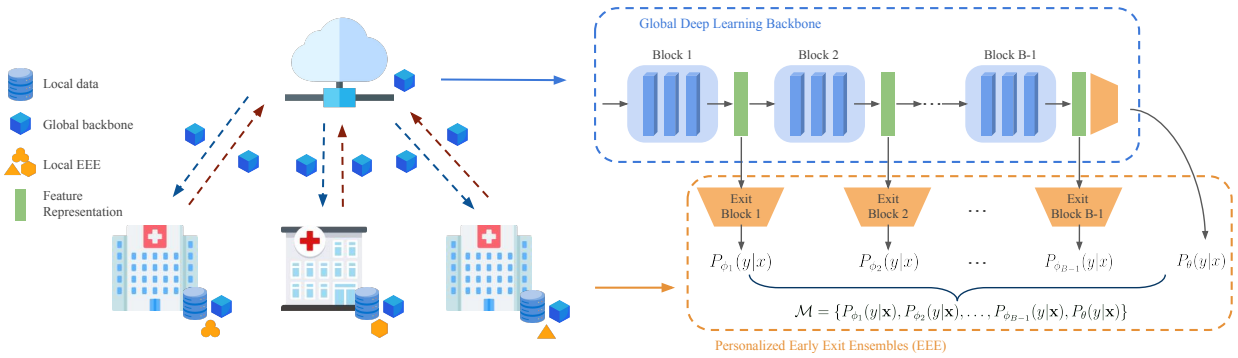


Figure 3: Design of the FedEE framework. The global backbone is trained using FL, while the light-weighted local early exit ensembles (EEE) remains personalized for uncertainty estimation.

work architecture, let's denote the neural network as θ and the entire model (or collection of models) as Φ.

3.1. Personalized FL Objective

Suppose there are K clients each with local data $D_k = \{x_i, y_i\}_{i=1}^{n_k}$. The standard FL objective is to learn a global model Φ that performs well for all the clients on average. However, the identical global model usually struggle to have desirable performance for every client under significant data 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 Φ_k. Then we consider the objective of personalized Federated Learning (pFL) (Li et al., 2021a; Tan et al., 2021),

$$\min_{\Phi_1, \Phi_2, \ldots, \Phi_K} F(\Phi) := \sum_{k=1}^{K} p_k \cdot F_k(\Phi_k), \tag{1}$$

where we optimize an individual model Φ_k for each client k that performs well on its local data D_k, $F_k(\cdot)$ is the local objective function for the k-th client weighted by p_k ($p_k \geq 0$ and $\sum_k p_k = 1$).

3.2. Uncertainty Measurement

For a given neural network model θ and data (x, y), confidence can be measured by calculating the Shannon entropy of the predictive distribution (e.g., softmax of the output),

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

where $\mathcal{H}$ represents the calculated entropy, $p_i = p(y = c_i|\mathbf{x}, \theta)$ represents the predicted probability of class i, and the summation is taken over C classes.

However, softmax entropy of a single deterministic model is often over-confident (Guo et al., 2017). To mitigate this issue, MC-Dropout, deep ensembles and early exit ensembles aim to inject randomness by sampling an ensemble of predictions during inference, $\mathcal{M} = \{P(y|\mathbf{x}, \theta_1), P(y|\mathbf{x}, \theta_2), \ldots, P(y|\mathbf{x}, \theta_{|\mathcal{M}|})\}$. Final predictions are obtained by their mean, and uncertainty can be measured by the predictive entropy by calculating p_i in Equation (2) using $p_i = \frac{1}{|\mathcal{M}|} \sum_{t=1}^{|\mathcal{M}|} p(y = c_i|x, \theta_t)$, where $|\mathcal{M}|$ is the size of the ensemble members.

3.3. Research Goals

In our study, we care about both the predictive accuracy and uncertainty estimates of the local models on respective local data. This is also the motivation of participating clients for an accurate and uncertainty-aware model. The goal of uncertainty-aware personalized federated learning is two-fold.

1. We want to learn through federated learning a personalized deep learning model Φ_k for each client k, which achieves high predictive accuracy on its local data $D_k = \{x_i, y_i\}_{i=1}^{n_k}$.
2. The personalized model Φ_k is expected to accurately assess the confidence in its predictions by quantifying the predictive entropy $\mathcal{H}(y|x, \Phi_k)$. Higher uncertainty should correspond to situations where the model lacks confidence in its predictions. We formulate two tasks, in line with existing literature (Combalia et al., 2020; Kompa et al., 2021), representing our goal of identifying potential misclassifications and enhancing the reliability of predictions.
   - **Misdiagnosis Detection.** We assess the model's ability to differentiate between correct and incorrect predictions, identifying cases that may require the attention of physicians. AUROC of misdiagnosis detection is defined as

$$\text{AUROC} = \frac{1}{N_{\text{M}} N_{\text{C}}} \sum_{i \in \text{M}} \sum_{j \in \text{C}} \mathbb{1}(\sigma_i > \sigma_j), \quad (3)$$

where M refers to the misdiagnoses and C refers to the correct diagnoses, $\mathbb{1}$ is the indicator function and σ_x is the quantified uncertainty of data sample x.

- **Selective Prediction.** We discard the most uncertain samples (leaving them for expert human review) from the test dataset and evaluate the prediction performance of the remaining data. The evaluation metric of selective prediction is defined as:

$$\text{P}_{SEL} = \text{P}(\{x \in X \mid \sigma_x \leq \sigma_h\}), \quad (4)$$

where P is the performance metric (e.g. accuracy), X is the dataset, σ_x is the quantified uncertainty of data sample x, and σ_h is the estimated uncertainty for threshold h, e.g. upper 40%.

4. Method

4.1. Overview

To achieve the aforementioned goals towards trustworthy health applications, we propose **FedEE**, an uncertainty-aware and efficient personalized FL framework. In response to the challenges posed by health data heterogeneity and the high computational overhead for decentralized uncertainty estimates, FedEE employs Early Exit Ensembles (EEE) to achieve both personalization and efficient uncertainty quantification. An overview of FedEE is presented in Figure 3: the left part illustrates the federated iterations, while the right side shows the learning process of the global model and local EEEs. We elaborate on the key components of FedEE below.

4.2. Early Exit Ensembles Model Structure

Effectively estimating uncertainty requires introducing randomness into deep learning models (Abdar et al., 2021). Since we want to achieve uncertainty-awareness while maintaining efficiency, leveraging diverse information from the model becomes important (Shen et al., 2023). Our approach leverages diverse feature representations from different layers of a single deep learning model. These features are passed through lightweight early exit blocks integrated into the backbone, creating an efficient ensemble.

Consider a neural network θ as a sequence of B blocks as depicted in the top blue dashed box in Figure 3, and $\theta = \cup_{i=1}^{B} \theta_i$ represents the union of each block's parameters (each block's parameters are mutually exclusive). In an early exit ensemble, we incorporate an exit block g_{ϕ_i} after each chosen exit i leveraging the intermediary output, each producing an prediction $P_{\phi_i}(y|\mathbf{x})$, as illustrated in the orange box in Figure 3. In this way, we get an ensemble of predictions,

$$\mathcal{M} = \{P_{\phi_1}(y|\mathbf{x}), P_{\phi_2}(y|\mathbf{x}), \ldots, P_{\phi_{B-1}}(y|\mathbf{x}), P_{\theta}(y|\mathbf{x})\}, \quad (5)$$

with size $|\mathcal{M}| = B$. The choice of exits is flexible with the specific task and architecture. The full model Φ constitutes of the backbone θ and the exits $\{\phi\}$.

During training, a weighted sum of each exit block's predictive loss is used to jointly train all the exits in the ensemble,

$$\mathcal{L} = \mathcal{L}(y, f_\theta(y|\mathbf{x})) + \Sigma_{i=1}^{B-1} \alpha_i \mathcal{L}(y, g_{\phi_1}(y|\mathbf{x})), \quad (6)$$

where $\alpha_i \in [0, 1]$ is a weight associated with the i-th exit, adjusting the importance of each exits. During inference, the final prediction is obtained by averaging the ensemble predictions.

Early exit ensembles achieve diversity by combining information from different layers of the backbone in a single forward pass, creating an implicit ensemble of models with varying depths. This approach is compatible with any multi-layer feed-forward network. In our work, we design our early exit blocks to be light-weight. Following Qendro et al. (2021), the i-th early exit meta block is defined as:

$$g_{\phi_i}(\mathbf{h}^{(i)}) = \mathbf{W}_2^{(i)} \sigma(\mathbf{W}_1^{(i)} f(\mathbf{h}^{(i)}) + \mathbf{b}_1^{(i)}) + \mathbf{b}_2^{(i)}, \quad (7)$$

where $\mathbf{W}$ and $\mathbf{b}$ are weights and bias matrices of linear layers, $f(\cdot)$ is a average pooling layer, $\sigma(\cdot)$ is an activation function such as RELU.

To address the potential impact of weaker representation power in earlier exits on accuracy, a conditional architecture is adopted. The number of features in each exit block is inversely proportional to the exit point so that earlier exits have access to a larger number of parameters to learn more complex relations between features.

4.3. Personalized Federated Training

Recognizing the potential of personalization in enhancing uncertainty estimation quality, which is pronounced in the example in Figure 1(*b*), we choose

Algorithm 1: FedEE. θ is the backbone model parameters, and ϕ refers to the EE blocks. D_k (size n_k) is client k's training data, E is the number of local epochs, and η is the learning rate.

Server executes:
Initialize global backbone model θ_0
for *each client* k **do**
  Randomly initialize local exit blocks ϕ_k
end
for *each round* $t = 0, 1, 2, \ldots$ **do**
  Send θ_t to all participating clients
  for *each client* $k \in S_t$ **do**
    $\theta_{t+1}^{k*} \leftarrow$ ClientUpdate()
  end
  $m_t \leftarrow \sum_{k \in S_t} n_k$
  $\theta_{t+1} \leftarrow \sum_{k \in S_t} \frac{n_k}{m_t} \theta_{t+1}^{k*}$ ▷ Server aggregation
end
Return each personalized model $\Phi_k \leftarrow \theta \cup \phi_k$

ClientUpdate():
Receive θ_t from server, fork $\theta_k \leftarrow \theta_t$
for *each local epoch* $e = 1, 2, \ldots E$ **do**
  $g_t^k = \nabla \mathcal{L}(D_k, \theta_k \cup \phi_k)$ ▷ Local objective
  $(\theta_k \cup \phi_k) \leftarrow (\theta_k \cup \phi_k) - \eta g_t^k$ ▷ Joint training
end
Return global model θ_k

to adopt a personalized approach in our federated uncertainty estimation. To achieve this, we introduce a novel personalized training strategy to the framework, where we train each client's early exit meta-models using only the local data, and plug them to the globally synchronized backbone deep learning model. The pseudo-code is detailed in Algorithm 1. The distinctive feature of this approach is the inclusion of local early exit blocks ϕ, enabling uncertainty awareness and personalization. The framework can accommodate other personalization and optimization techniques by modifying the local objective or aggregation function.

4.4. Analysis of Computation and Communication Cost

Table 1 compares costs between FedEE and existing methods. Deep ensembles are the most resource-intensive, requring training $|\mathcal{M}|$ independent models. This cost is exacerbated in FL, spreading the computation cost across every client at each training round and local epoch, and leading to increased network communication overhead. MC-Dropout, while not introducing costs during federated training, involves multiple forward passes during inference, which is also computationally expensive. In contrast, FedEE uses a single model with one forward pass, with memory and computational overhead only from the exit block parameters $\phi = \cup_{i=1}^{|\mathcal{M}|-1} \phi_i$. Since generally $|\phi| \ll |\theta|$, memory and computation cost are reduced by leveraging shared weights between the ensembles.

Table 1: FL computational and communicational costs of MC-Dropout, deep ensembles and FedEE. $|\mathcal{M}|$ is the ensemble size, F and τ denotes the number of FLOPs (floating-point operations) during inference and training, and $\mathcal{R}$ is the number of communication rounds.

Method	Size	FLOPs	Comp.	Comm.
Backbone	$\|\theta\|$	F	τ	$\|\theta\| * \mathcal{R}$
MC-Dropout	$\|\theta\|$	$F * \|\mathcal{M}\|$	τ	$\|\theta\| * \mathcal{R}$
Deep Ensembles	$\|\theta\| * \|\mathcal{M}\|$	$F * \|\mathcal{M}\|$	$\tau * \|\mathcal{M}\|$	$\|\theta\| * \mathcal{R} * \|\mathcal{M}\|$
FedEE	$\|\theta\| + \|\phi\|$	$F + F_\phi$	$\tau + \tau_\phi$	$\|\theta\| * \mathcal{R}$

5. Experiments

5.1. Datasets

For a comprehensive and realistic evaluation, we utilize three diverse real-world health datasets with varying sample sizes, input modalities, and levels of data heterogeneity. Table. 2 presents key characteristics of these datasets, which are naturally partitioned from different sources. The datasets exhibit universal data heterogeneity among clients, including variations in sample numbers, label distributions, patient demographics and data collection technologies. Further details are provided in Appendix A.

5.2. Experimental Setup

To our knowledge, we are the first to study uncertainty quantification in the FL setting using heterogeneous health data. Therefore, we set up our evaluation framework by referencing research on centralized uncertainty estimation (Combalia et al., 2020; Kompa et al., 2021) and personalized federated learning (Chen et al., 2022; Li et al., 2021a).

Uncertainty estimation baselines. We implemented three widely-used uncertainty estimation methods from centralized deep learning in the FL setting for comparison. The three baselines are the Vanilla Softmax entropy (**Backbone**), which utilizes only one inference drawn from the backbone architec-

Table 2: Summary of datasets and FL partition.

Dataset	PAMAP2	ISIC2019	PhysioNet-2016
Modality	time series	image	audio
Task	human activity recognition	melanoma class prediction	heart sound classification
# Samples	2,869	23,247	13,015
# Clients	8	6	6
# Classes	8	8	2
Train Partition	173, 171, 179, 181, 176, 179, 188, 185	9930, 3163, 2691, 1807, 655, 351	1554, 294, 174, 82, 5276, 427
Test Partition	174, 172, 179, 182, 177, 180, 188, 185	2483, 791, 672, 452, 164, 88	1036, 196, 117, 56, 3518, 285
Input Dimension	1000×3	$200 \times 200 \times 3$	$101 \times 99 \times 1$

ture; MC-Dropout (**MCDrop**), which keeps dropout layers open for inference ($|\mathcal{M}| = 5$); and Deep Ensembles (**Deep Ens.**), which trains multiple global models ($|\mathcal{M}| = 5$) with different weight initializations. Unless otherwise stated, our default federated aggregation strategy is FedAvg.

Evaluation metrics. We report three metrics for each method, one for classification results (**Acc.**) and two for the quality of the uncertainty estimates measured by the performance of misdiagnosis detection (**Mis. Det.**) and selective prediction (**Sel. Pred.**). For misclassification detection, we use the area under the receiver operating curve (AUROC) for classification based on the predictive entropy. For selective prediction, we observe consistent performance across different thresholds, and report at a 40% threshold, showing clear accuracy improvements after abstention and effectively distinguishing different methods. For each of the metric, we report the weighted average performance across clients, with a discussion of the averaging scheme in the Appendix.

5.3. Classification and Uncertainty Estimation Performance

Following the setup, we now report the main experimental results. The classification and uncertainty estimation metrics for Backbone, MCDropout, Deep Ensembles and FedEE are summarized in Table 3. The results showcase that FedEE consistently outperforms all baselines across metrics and datasets. In terms of uncertainty estimation, it achieves a relative improvement of up to 15% in misdiagnosis detection and 12% in selective prediction compared to the backbone. Moreover, it surpasses the more expensive baselines by 2% to 9%. Additionally, FedEE shows an accuracy improvement of 3% to 8%, likely benefiting from the personalized training and diverse feature utilization.

In real-world applications, the optimal number of uncertain data points to remove depends on the desired accuracy level and the availability of expert resources to review potentially uncertain points. Hence, in addition to the selective prediction performance at a fixed threshold in Table 3, we further examine how predictive accuracy evolves with an increasing percentage of discarded samples. As illustrated in Figure 4, FedEE consistently outperforms other methods, especially at smaller thresholds. Notably, FedEE achieves the same accuracy by discarding only 20% of samples, while other baselines require a drop of at least 40%, reducing the workload for additional physician examinations. The steeper slope indicates the higher quality of estimated uncertainty. Not only are wrong predictions reduced, but they are also more easily identifiable, as evidenced by a higher AUROC in misdiagnosis detection.

5.4. Improvement in Efficiency

Having shown FedEE's superior performance in classification and uncertainty estimation, we now delve into a detailed analysis of the induced overhead. Though deep ensembles also show good performance, FedEE brings notable efficiency improvement to it, reducing computation and communication cost by a factor of $|\mathcal{M}|$, where $|\mathcal{M}| = 5$ in our case. Having analyzed theoretically in Table 1, we further compare the actual estimated costs in the experiments.

Table 4 summarizes the memory usage (size of model) and FLOPs during inference. As expected, deep ensembles have the highest memory footprint since it consists of 5 independently trained models. The $5\times$ increase in size for deep ensembles translates to $5\times$ communication cost during each FL round,

Table 3: Performance for classification, misdiagnosis detection, and selective prediction. Best method in bold, and relative improvement of FedEE compared to backbone and the best baseline shown in Δ.

Method	PAMAP2			ISIC-2019			PhysioNet-2016		
	Acc. (↑)	Mis. Det. (↑)	Sel. Pred. (↑)	Acc. (↑)	Mis. Det. (↑)	Sel. Pred. (↑)	Acc. (↑)	Mis. Det. (↑)	Sel. Pred. (↑)
Backbone	0.780 ± 0.01	0.769 ± 0.02	0.866 ± 0.02	0.710 ± 0.02	0.804 ± 0.01	0.827 ± 0.03	0.856 ± 0.01	0.772 ± 0.07	0.887 ± 0.01
MCDrop	0.782 ± 0.02	0.817 ± 0.01	0.894 ± 0.02	0.707 ± 0.02	0.790 ± 0.01	0.815 ± 0.02	0.842 ± 0.01	0.847 ± 0.03	0.880 ± 0.01
Deep Ens.	0.787 ± 0.01	0.788 ± 0.01	0.877 ± 0.01	0.725 ± 0.00	0.804 ± 0.00	0.847 ± 0.01	0.855 ± 0.01	0.871 ± 0.01	0.889 ± 0.01
FedEE	**0.840 ± 0.01**	**0.856 ± 0.01**	**0.942 ± 0.01**	**0.748 ± 0.01**	**0.822 ± 0.01**	**0.924 ± 0.01**	**0.879 ± 0.02**	**0.885 ± 0.02**	**0.945 ± 0.01**
Δ (%)	(7.7/6.7)	(11.3/4.8)	(8.8/5.4)	(5.4/3.2)	(2.2/2.2)	(11.7/9.1)	(2.7/2.8)	(14.6/1.6)	(6.5/6.3)

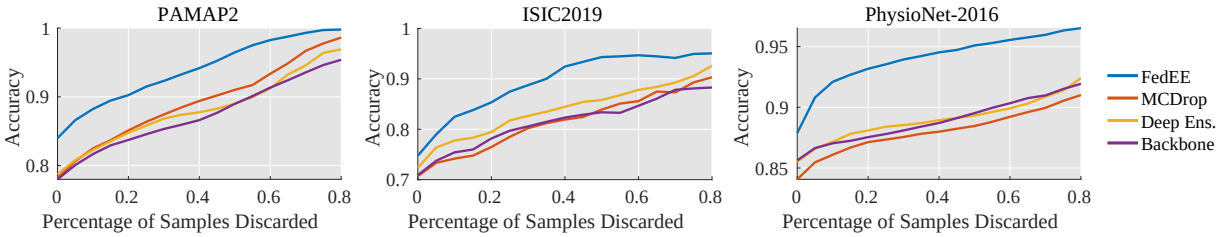


Figure 4: Evolution of the predictive performance as the most uncertain samples are removed.

Table 4: Efficiency evaluation of memory (Millions) and inference FLOPs (Giga).

		Backbone	MCDrop	Deep Ens.	FedEE
PAMAP2 (3-layer CNN)	Size (↓)	0.953	0.953	4.765	0.963
	FLOPs (↓)	0.836	4.180	4.180	0.836
ISIC2019 (EfficientNet-b0)	Size (↓)	4.018	4.018	20.090	4.725
	FLOPs (↓)	20.374	101.870	101.870	20.422
PhysioNet-2016 (ResNet18)	Size (↓)	11.171	11.171	55.855	11.753
	FLOPs (↓)	27.657	138.285	138.285	27.700

and 5× local training time. This gap increases with each round in training, with the estimated computational and communicational cost illustrated in Figure 10 in the Appendix. MC-Dropout is the most memory-efficient but requires 5× the FLOPs for uncertainty estimation during inference and has lower performance in classification and uncertainty estimation. In contrast, FedEE introduces only a small additional memory overhead while significantly improving performance.

Overall, FedEE strikes a better trade-off between performance and costs, inducing minimal overhead in memory demand, training time, server-client communication and inference time, with comparable if not better performance compared to the baselines.

5.5. Additional Results and Discussion

Having demonstrated FedEE's effectiveness and superiority in tackling the challenges in federated uncertainty estimation, we conduct additional experiments to further explore FedEE's capabilities.

Trade-off between classification performance and expert workload. To maximize the benefits of selective prediction, we delve deeper into investigating the number of samples requiring expert consultation to reach certainty classification performance. For instance, Figure 5 illustrates the percentage of data points that need to be referred if we aim for an overall accuracy of 90% or 95% (assuming that all cases referred to experts will be correctly diagnosed). (The PAMAP2 dataset contains a total of 1,434 test samples, ISIC2019 has 4,650, and PhysioNet-2016 has 5,208.) Notably, FedEE significantly outperforms the baselines at both performance level by requiring fewer samples to achieve the target accuracy, less than 30% for 90% accuracy and less than 40% for 95% accuracy, thereby imposing a lower burden on clinicians.

Compatibility with other FL techniques. We implement FedEE combined with **FedProx**, and personalized FL methods **FedBN**, **FedAP**, along with **Fine-tuning (FT)**. The results on PAMAP2 are detailed in Table 5. FedEE demonstrates compatibility with other personalization methods, where up to 17% improvement in misclassification detection and 16% in selective prediction can be achieved. It consistently outperforms vanilla softmax entropy and MC-Drop across datasets and combined strategies. Remarkably, FedEE surpasses the five times more resource-intensive deep ensembles in over half of the cases, offering *the best or comparable performance in uncertainty estimation with significantly higher efficiency.*

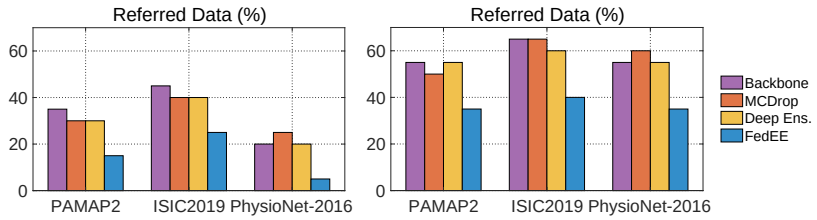


Figure 5: Percentage of data samples required for referring to doctors in order to achieve accuracies of 90% (a) and 95% (b).

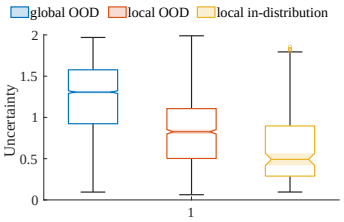


Figure 6: Uncertainty estimates of local ID, local OOD and global OOD.

Table 5: Performance results when combined with other strategies. FedEE ranks first or second in all cases for uncertainty estimation. More results in App. C.

Strategy	Method	PAMAP2		
		Acc. (↑)	Mis. Det. (↑)	Sel. Pred. (↑)
FedProx	Backbone	0.775 ± 0.01	0.771 ± 0.01	0.868 ± 0.02
	MCDrop	0.778 ± 0.01	0.812 ± 0.02	0.890 ± 0.01
	Deep Ens.	0.782 ± 0.00	0.775 ± 0.01	0.866 ± 0.01
	FedEE	**0.842 ± 0.01**	**0.848 ± 0.01**	**0.935 ± 0.01**
FedAvg+FT	Backbone	0.877 ± 0.01	0.878 ± 0.01	0.980 ± 0.01
	MCDrop	0.870 ± 0.01	0.895 ± 0.01	0.986 ± 0.01
	Deep Ens.	0.894 ± 0.00	**0.907 ± 0.00**	0.991 ± 0.00
	FedEE	**0.908 ± 0.01**	0.896 ± 0.01	**0.993 ± 0.00**
FedProx+FT	Backbone	0.869 ± 0.00	0.860 ± 0.02	0.971 ± 0.00
	MCDrop	0.866 ± 0.00	0.888 ± 0.01	0.981 ± 0.00
	Deep Ens.	0.892 ± 0.00	**0.902 ± 0.00**	0.991 ± 0.00
	FedEE	**0.909 ± 0.00**	0.895 ± 0.01	**0.992 ± 0.00**
FedBN(+FT)	Backbone	0.870 ± 0.01	0.870 ± 0.02	0.973 ± 0.01
	MCDrop	0.867 ± 0.01	0.890 ± 0.01	0.981 ± 0.01
	Deep Ens.	0.897 ± 0.00	**0.902 ± 0.00**	**0.994 ± 0.00**
	FedEE	**0.910 ± 0.01**	0.896 ± 0.01	0.992 ± 0.00
FedAP(+FT)	Backbone	0.873 ± 0.01	0.874 ± 0.02	0.977 ± 0.01
	MCDrop	0.868 ± 0.01	0.885 ± 0.01	0.980 ± 0.01
	Deep Ens.	0.897 ± 0.00	**0.907 ± 0.01**	0.992 ± 0.00
	FedEE	**0.921 ± 0.01**	0.906 ± 0.01	**0.995 ± 0.00**

Out-of-distribution detection. In addition to identifying difficult samples from the same distribution as the local training data, another key application of uncertainty estimation is out-of-distribution (OOD) detection. We want to investigate how our models respond when confronted with data from a distribution completely outside the training data. Here we consider three data types: local in-distribution, local OOD (from another client's distribution), and global OOD (entirely new data). Using human activity recognition as an example, we evaluate FedEE's effectiveness in OOD detection. We use local test data as in-distribution (ID), other clients' test data as local OOD, and another data source, Opportunity dataset, as global OOD. FedEE generates gradually higher uncertainty estimates for the two types of OOD data (Figure 6), achieving a strong AUROC of 0.825 for global OOD detection.

6. Conclusions and Future Work

This paper pioneers uncertainty estimation in decentralized health applications, addressing key challenges of data heterogeneity and high computational cost with our novel FL framework, FedEE. Extensive experiments on three real-world datasets show FedEE's *competitive performance* and *significant efficiency improvements*. Our study contributes to trustworthy deep learning in healthcare by integrating uncertainty quantification and FL in a unified framework. This takes a significant step toward translating deep learning research into practical clinical applications.

Although our method already greatly reduces the workload on healthcare professionals using selective prediction, the number of samples needed for referral is still quite substantial in order to achieve a high accuracy (e.g. 95%). This could potentially be improved when there are more hospitals joining and more data available.

Due to FedEE's high efficiency, we anticipate widespread use of our methodology in the era of large-scale models or foundational models as the backbone for healthcare applications, where uncertainty estimation is crucial and efficiency is a bottleneck. Another intriguing future direction could be addressing system heterogeneity by enabling clients to retain only a subset of exit blocks and partial backbone models. Clients with varying computation resources can adapt different portions of the model, tailoring the framework to their specific capabilities.

Acknowledgments

This work was supported by ERC Project 833296 (EAR), EPSRC RELOAD, Cambridge Trusts, and Nokia Bell Labs through a donation.

References

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021.

Laith Alzubaidi, Jinglan Zhang, Amjad J Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, José Santamaría, Mohammed A Fadhel, Muthana Al-Amidie, and Laith Farhan. Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions. *Journal of big Data*, 8:1–74, 2021.

Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, et al. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 401–413, 2021.

Alexander Campbell, Lorena Qendro, Pietro Liò, and Cecilia Mascolo. Robust and efficient uncertainty aware biosignal classification via early exit ensembles. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3998–4002. IEEE, 2022.

Daoyuan Chen, Dawei Gao, Weirui Kuang, Yaliang Li, and Bolin Ding. pfl-bench: A comprehensive benchmark for personalized federated learning. *Advances in Neural Information Processing Systems*, 35:9344–9360, 2022.

Hong-You Chen and Wei-Lun Chao. Fedbe: Making bayesian model ensemble applicable to federated learning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=dgtpE6gKjHn.

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. 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*, pages 168–172, 2018.

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

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

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

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

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

José Miguel Hernández-Lobato and Ryan Adams. Probabilistic backpropagation for scalable learning of bayesian neural networks. In *International conference on machine learning*, pages 1861–1869. PMLR, 2015.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances

and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2): 1–210, 2021.

Christopher J Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine*, 17:1–9, 2019.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

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

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

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

Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning*, pages 6357–6368. PMLR, 2021a.

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

Florian Linsner, Linara Adilova, Sina Däubener, Michael Kamp, and Asja Fischer. Approaches to uncertainty quantification in federated deep learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 128–145. Springer, 2021.

Chengyu Liu, David Springer, Qiao Li, Benjamin Moody, Ricardo Abad Juan, Francisco J Chorro, Francisco Castells, José Millet Roig, Ikaro Silva, Alistair EW Johnson, et al. An open access database for the evaluation of heart sound algorithms. *Physiological measurement*, 37(12):2181, 2016.

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

David JC MacKay. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.

Yishay Mansour, Mehryar Mohri, Jae Ro, and Ananda Theertha Suresh. Three approaches for personalization with applications to federated learning. *arXiv preprint arXiv:2002.10619*, 2020.

Tanachat Nilanon, Jiayu Yao, Junheng Hao, Sanjay Purushotham, and Yan Liu. Normal/abnormal heart sound recordings classification using convolutional neural network. In *2016 computing in cardiology conference (CinC)*, pages 585–588. IEEE, 2016.

Lorena Qendro, Alexander Campbell, Pietro Lio, and Cecilia Mascolo. Early exit ensembles for uncertainty quantification. In *Machine Learning for Health*, pages 181–195. PMLR, 2021.

Wanyong Qiu, Kun Qian, Zhihua Wang, Yi Chang, Zhihao Bao, Bin Hu, Björn W Schuller, and Yoshiharu Yamamoto. A federated learning paradigm for heart sound classification. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1045–1048. IEEE, 2022.

Xinchi Qiu, Titouan Parcollet, Javier Fernandez-Marques, Pedro PB Gusmao, Yan Gao, Daniel J Beutel, Taner Topal, Akhil Mathur, and Nicholas D Lane. A first look into the carbon footprint of federated learning. *Journal of Machine Learning Research*, 24(129):1–23, 2023.

Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečnỳ, Sanjiv Kumar, and H Brendan McMahan. Adaptive federated optimization. *arXiv preprint arXiv:2003.00295*, 2020.

Attila Reiss and Didier Stricker. Introducing a new benchmarked dataset for activity monitoring. In

2012 16th international symposium on wearable computers, pages 108–109. IEEE, 2012.

Daniel Roggen, Alberto Calatroni, Mirco Rossi, Thomas Holleczek, Kilian Förster, Gerhard Tröster, Paul Lukowicz, David Bannach, Gerald Pirkl, Alois Ferscha, et al. Collecting complex activity datasets in highly rich networked sensor environments. In *2010 Seventh international conference on networked sensing systems (INSS)*, pages 233–240. IEEE, 2010.

Maohao Shen, Yuheng Bu, Prasanna Sattigeri, Soumya Ghosh, Subhro Das, and Gregory Wornell. Post-hoc uncertainty learning using a dirichlet meta-model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 9772–9781, 2023.

Naichen Shi, Fan Lai, Raed Al Kontar, and Mosharaf Chowdhury. Fed-ensemble: Improving generalization through model ensembling in federated learning. *arXiv preprint arXiv:2107.10663*, 2021.

Alysa Ziying Tan, Han Yu, Li zhen Cui, and Qiang Yang. Towards personalized federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

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

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. Flamby: Datasets and benchmarks for cross-silo federated learning in realistic healthcare settings. *arXiv preprint arXiv:2210.04620*, 2022.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, pages 1–9, 2018.

Hao Wang, Zakhary Kaplan, Di Niu, and Baochun Li. Optimizing federated learning on non-iid data with reinforcement learning. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*, pages 1698–1707. IEEE, 2020a.

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

Tong Xia, Jing Han, Abhirup Ghosh, and Cecilia Mascolo. Cross-device federated learning for mobile health diagnostics: A first study on covid-19 detection. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

Tong Xia, Ting Dang, Jing Han, Lorena Qendro, and Cecilia Mascolo. Uncertainty-aware health diagnostics via class-balanced evidential deep learning. *IEEE Journal of Biomedical and Health Informatics*, 2024a.

Tong Xia, Abhirup Ghosh, Xinchi Qiu, and Cecilia Mascolo. Flea: Addressing data scarcity and label skew in federated learning via privacy-preserving feature augmentation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3484–3494, 2024b.

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

Yuwei Zhang, Tong Xia, Abhirup Ghosh, and Cecilia Mascolo. Uncertainty quantification in federated learning for heterogeneous health data. In *International Workshop on Federated Learning for Distributed Data Mining*, 2023.

Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.

Junyi Zhu, Xingchen Ma, and Matthew B Blaschko. Confidence-aware personalized federated learning via variational expectation maximization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24542–24551, 2023.

Appendix A. Dataset Description

PAMAP2. The PAMAP2 dataset Reiss and Stricker (2012) 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. (2022). After preprocessing, the data contain 8 activity classes performed by 8 subjects, each acting as a client.

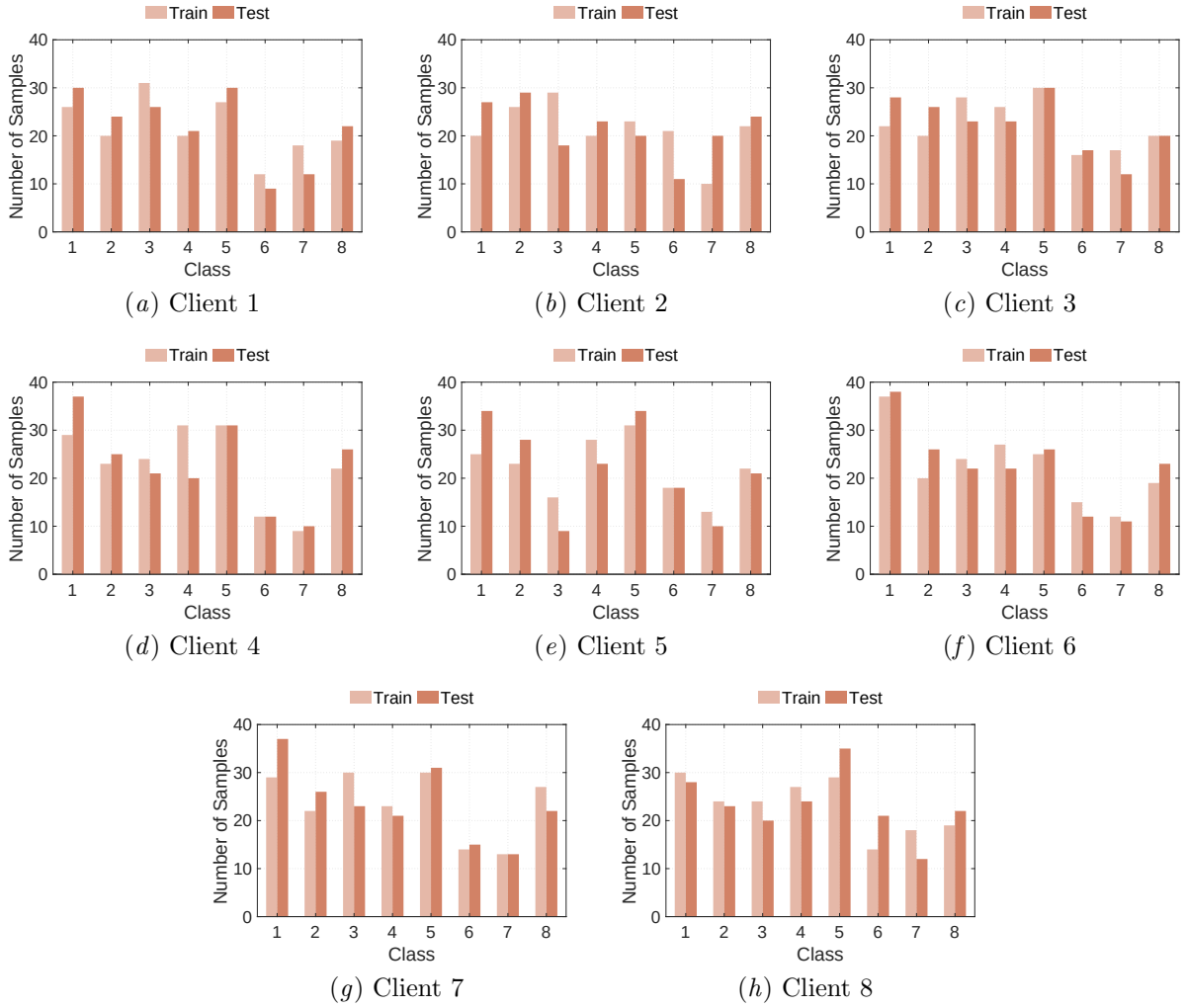
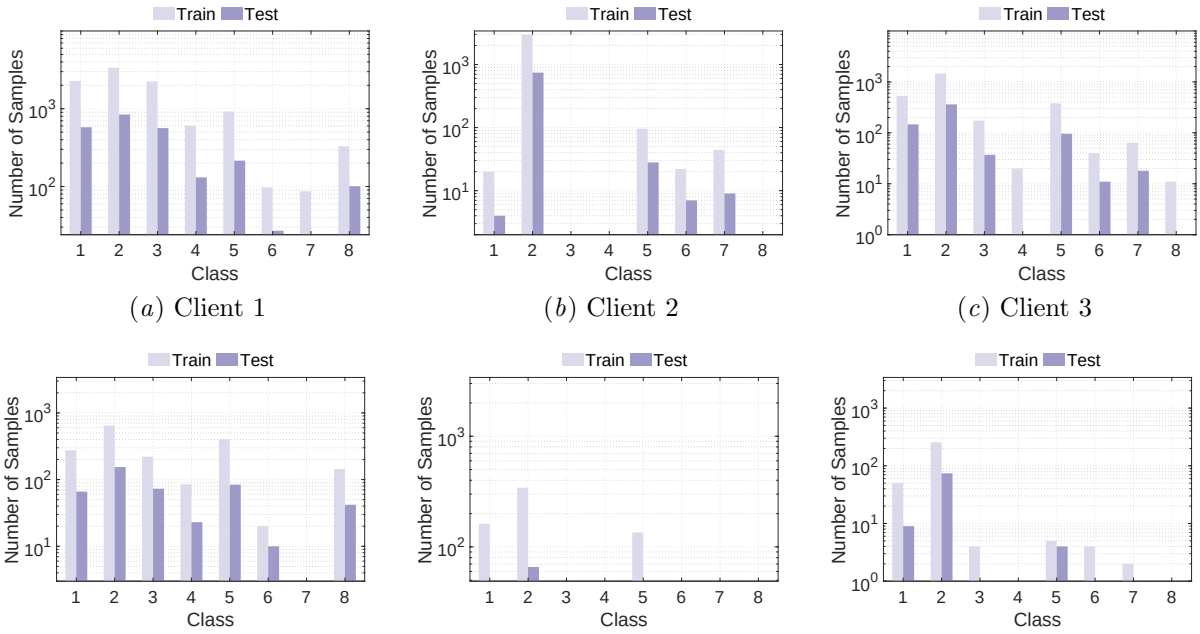


Figure 7: Label distribution of the 8 clients in PAMAP2 dataset.

ISIC2019. The ISIC2019 datasets (Tschandl et al., 2018; Codella et al., 2018; Combalia et al., 2019) 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. The baseline model is an EfficientNet-B0 (Tan and Le, 2019) pretrained on ImageNet following the FLamby benchmark (Terrail et al., 2022), along with preprocessing and data augmentation steps.

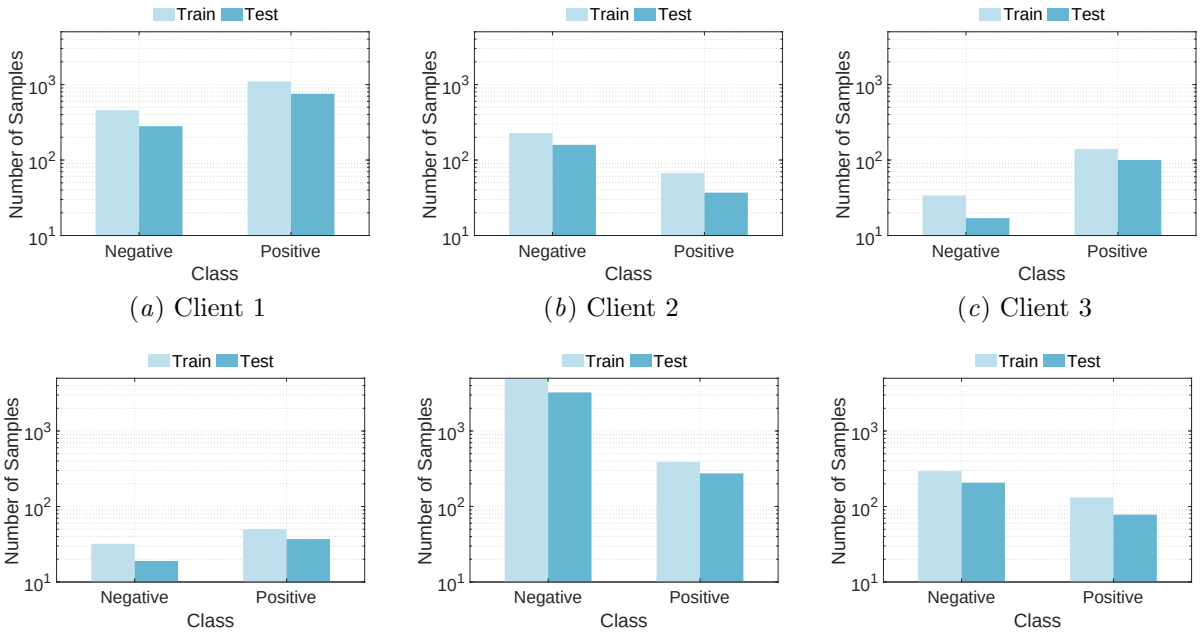
PhysioNet-2016. The PhysioNet-2016 dataset contains heart sound data from the PhysioNet/CinC Challenge 2016 (Liu et al., 2016; Goldberger et al., 2000), consisting of 3240 PCG heart sound recordings



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

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

Figure 8: Label distribution of the 6 clients in ISIC-2019 dataset in log-scale.



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

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

Figure 9: Label distribution of the 6 clients in PhysioNet-2016 dataset in log-scale.

from six independent databases sourced from multiple research groups around the world. The length of recording varies from several seconds to several minutes, so we segment them into 5-second segments following previous work using the same dataset (Nilanon et al., 2016; Qiu et al., 2022) with supporting assumption validated by practitioners (Nilanon et al., 2016). We then extract spectrograms from the audio segments as time-frequency feature input to a ResNet-18 model.

The label distributions of the three datasets are concluded in Fig. 7, Fig. 8 and Fig. 9.

In exploring OOD detection, we utilize an additional ***Opportunity dataset*** (Roggen et al., 2010), which consist of 3.9K samples of four subjects performing four activity classes. We follow the preprocessing of precious work (Yuan et al., 2022).

Appendix B. Implementation Details

Table. 6 presents the chosen setup for training and evaluation. The optimizer used in all experiments is Adam (Kingma and Ba, 2015). 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 (Terrail et al., 2022), and is same for all the strategies. The number of rounds $\mathcal{R}$ is calculated as

$$\mathcal{R} = n_{epoches}^{pooled} \cdot \lfloor N/K/B/E \rfloor, \tag{8}$$

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.

Table 6: Implementation details.

Dataset	**ISIC2019**	**PAMAP2**	**PhysioNet-2016**
Model	EfficientNet-B0	CNN	ResNet18
Batch Size	64	32	64
Learning Rate	0.005	0.0003	0.0001
# Local Iters	20	20	20
# Rounds	47	10	32
Metric	Balanced Accuracy	Macro F1-score	Accuracy
Loss	Weighted Focal Loss	CE Loss	BCE Loss

Our hyperparameters in the implementation details are selected with reference to previous research and the common practices of each application in a centralized setting. We did not observe significant sensitivity in the hyperparameters. For the ISIC2019 dataset, we adhere to the FLamby benchmark for all hyperparameter and loss function selections, including a weighted focal loss for addressing class imbalance. For the other two datasets, we use simpler standard loss functions of (binary) cross entropy loss and the learning rate and batch size are common choices depending on the application. Below are FedEE's results on PAMAP2 with varying learning rates and batch sizes. Performance remained stable with less than 1% fluctuation.

Table 7: Accuracy on PAMAP2 with learning rate ranging from [0.0001, 0.0003, 0.0005] and batch size ranging from [16, 32, 64].

	0.0001	**0.0003**	**0.0005**	**0.001**
16	0.848 ± 0.007	0.844 ± 0.008	0.847 ± 0.010	0.841 ± 0.005
32	0.848 ± 0.005	0.848 ± 0.011	0.845 ± 0.008	0.847 ± 0.010
64	0.843 ± 0.001	0.840 ± 0.006	0.851 ± 0.007	0.845 ± 0.011

Table 8: Misclassification detection performance (AUROC) on PAMAP2 with learning rate ranging from [0.0001, 0.0003, 0.0005] and batch size ranging from [16, 32, 64].

	0.0001	**0.0003**	**0.0005**	**0.001**
16	0.859 ± 0.013	0.854 ± 0.006	0.860 ± 0.007	0.849 ± 0.014
32	0.852 ± 0.019	0.863 ± 0.005	0.859 ± 0.013	0.846 ± 0.009
64	0.848 ± 0.012	0.863 ± 0.005	0.857 ± 0.013	0.853 ± 0.017

Table 9: Selective prediction performance on PAMAP2 dataset with learning rate ranging from [0.0001, 0.0003, 0.0005] and batch size ranging from [16, 32, 64].

	0.0001	**0.0003**	**0.0005**	**0.001**
16	0.948 ± 0.014	0.944 ± 0.009	0.948 ± 0.016	0.939 ± 0.013
32	0.948 ± 0.012	0.951 ± 0.004	0.942 ± 0.013	0.942 ± 0.011
64	0.938 ± 0.006	0.945 ± 0.009	0.949 ± 0.014	0.945 ± 0.014

We train on one NVIDIA A100 GPU, and set number of forward passes T for MC-Dropout and number of models M for deep ensembles the same as number of exits for a fair comparison. 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.

Appendix C. Additional Results

Compatibility with other FL techniques. Table 10 presents the performance of FedEE and baselines in all the datasets and all the FL and pFL strategies.

Ablation study. To understand the benefit of personalization in the proposed method, we also conducted a comparison with the centralized version of Early Exit Ensembles (FedEE without personalization). The results are shown in Table 11.

Computational Costs and Carbon Emission. Figure 10 shows the comparison of estimated communication cost, computation cost and carbon emission of FedEE and deep ensembles using all three datasets. We can see that the gap becomes larger as the training round increases. Note that the minimal difference between FedEE and the backbone (and MC-Dropout) is so negligible that the curves practically overlap.

The carbon emission is calculated following Qiu et al. (Qiu et al., 2023), presented in Table 12. We repeatedly query the NVIDIA System Management Interface (NVIDIA-smi) to sample the GPU power consumption and report the average over all processed samples while training. The total training energy consumption of K clients with hardware power e for a total of R rounds is calculated as:

$$T(e, K, R) = \sum_{j=1}^{R} \sum_{k=1}^{K} t_k e_k, \tag{9}$$

where t_i is the wall clock time per round and e_k the power of client k.

The communication carbon is estimated as:

$$C(e, K, R) = \sum_{j=1}^{R} \sum_{k=1}^{K} S(\frac{1}{D} + \frac{1}{U})(e_r + e_{k,idle}) \tag{10}$$

Table 10: Complete performance results on three datasets, including predictive accuracy (Acc.), AUROC for misdiagnosis detection (Mis. Det.) and predictive accuracy for selective prediction (Sel. Pred.). We can see that FedEE achieves best or second performance in all cases for uncertainty estimation.

Strategy	Method	PAMAP2			ISIC-2019			PhysioNet-2016		
		Acc. (↑)	Mis. Det. (↑)	Sel. Pred. (↑)	Acc. (↑)	Mis. Det. (↑)	Sel. Pred. (↑)	Acc. (↑)	Mis. Det. (↑)	Sel. Pred. (↑)
FedAvg	Backbone	0.780 ± 0.01	0.769 ± 0.02	0.866 ± 0.02	0.710 ± 0.02	0.804 ± 0.01	0.827 ± 0.03	0.856 ± 0.01	0.772 ± 0.07	0.887 ± 0.01
	MCDrop	0.782 ± 0.02	0.817 ± 0.01	0.894 ± 0.02	0.707 ± 0.02	0.790 ± 0.01	0.815 ± 0.02	0.842 ± 0.01	0.847 ± 0.03	0.880 ± 0.01
	Deep Ens.	0.787 ± 0.01	0.788 ± 0.01	0.877 ± 0.01	0.725 ± 0.00	0.804 ± 0.00	0.847 ± 0.01	0.855 ± 0.01	0.871 ± 0.01	0.889 ± 0.01
	FedEE	**0.840 ± 0.01**	**0.856 ± 0.01**	**0.942 ± 0.01**	**0.748 ± 0.01**	**0.822 ± 0.01**	**0.924 ± 0.01**	**0.879 ± 0.02**	**0.885 ± 0.02**	**0.945 ± 0.01**
FedProx	Backbone	0.775 ± 0.01	0.771 ± 0.01	0.868 ± 0.02	0.754 ± 0.01	0.831 ± 0.01	0.895 ± 0.01	0.689 ± 0.15	0.614 ± 0.09	0.728 ± 0.18
	MCDrop	0.778 ± 0.01	0.812 ± 0.02	0.890 ± 0.01	0.754 ± 0.01	0.801 ± 0.01	0.876 ± 0.01	0.728 ± 0.17	0.572 ± 0.14	0.740 ± 0.21
	Deep Ens.	0.782 ± 0.00	0.775 ± 0.01	0.866 ± 0.01	**0.771 ± 0.00**	**0.835 ± 0.00**	0.916 ± 0.00	0.796 ± 0.03	0.522 ± 0.05	0.783 ± 0.07
	FedEE	**0.842 ± 0.01**	**0.848 ± 0.01**	**0.935 ± 0.01**	*0.748 ± 0.01*	0.830 ± 0.01	**0.919 ± 0.02**	**0.892 ± 0.03**	**0.857 ± 0.03**	**0.940 ± 0.02**
FedAvg+FT	Backbone	0.877 ± 0.01	0.878 ± 0.01	0.980 ± 0.01	0.760 ± 0.03	0.839 ± 0.01	0.920 ± 0.02	0.920 ± 0.00	0.843 ± 0.04	0.950 ± 0.00
	MCDrop	0.870 ± 0.01	0.895 ± 0.01	0.986 ± 0.01	0.756 ± 0.02	0.831 ± 0.01	0.867 ± 0.02	0.929 ± 0.00	0.850 ± 0.04	0.956 ± 0.00
	Deep Ens.	0.894 ± 0.00	**0.907 ± 0.00**	0.991 ± 0.00	**0.783 ± 0.01**	0.866 ± 0.01	0.953 ± 0.02	0.930 ± 0.00	0.903 ± 0.00	0.961 ± 0.00
	FedEE	**0.908 ± 0.01**	0.896 ± 0.01	**0.993 ± 0.00**	0.781 ± 0.00	**0.875 ± 0.01**	**0.957 ± 0.00**	**0.936 ± 0.00**	**0.917 ± 0.00**	**0.969 ± 0.00**
FedProx+FT	Backbone	0.869 ± 0.00	0.860 ± 0.02	0.971 ± 0.00	0.757 ± 0.02	0.841 ± 0.01	0.908 ± 0.05	0.920 ± 0.00	0.881 ± 0.02	0.949 ± 0.01
	MCDrop	0.866 ± 0.00	0.888 ± 0.01	0.981 ± 0.00	0.750 ± 0.02	0.830 ± 0.02	0.891 ± 0.03	0.927 ± 0.00	0.887 ± 0.02	0.956 ± 0.00
	Deep Ens.	0.892 ± 0.00	**0.902 ± 0.00**	0.991 ± 0.00	**0.796 ± 0.01**	0.869 ± 0.01	**0.962 ± 0.00**	0.934 ± 0.00	0.907 ± 0.00	0.962 ± 0.00
	FedEE	**0.909 ± 0.00**	0.895 ± 0.01	**0.992 ± 0.00**	0.781 ± 0.01	**0.876 ± 0.00**	0.955 ± 0.01	**0.940 ± 0.01**	**0.912 ± 0.01**	**0.972 ± 0.00**
FedBN(+FT)	Backbone	0.870 ± 0.01	0.870 ± 0.02	0.973 ± 0.01	0.756 ± 0.05	0.817 ± 0.03	0.898 ± 0.06	0.922 ± 0.00	0.835 ± 0.06	0.951 ± 0.00
	MCDrop	0.867 ± 0.01	0.890 ± 0.01	0.981 ± 0.01	0.753 ± 0.04	0.774 ± 0.02	0.872 ± 0.06	0.927 ± 0.00	0.804 ± 0.06	0.956 ± 0.00
	Deep Ens.	0.897 ± 0.00	**0.902 ± 0.00**	**0.994 ± 0.00**	**0.808 ± 0.01**	0.822 ± 0.01	**0.951 ± 0.00**	0.933 ± 0.00	0.898 ± 0.01	0.960 ± 0.00
	FedEE	**0.910 ± 0.01**	0.896 ± 0.01	0.992 ± 0.00	0.781 ± 0.00	**0.875 ± 0.01**	0.949 ± 0.02	**0.945 ± 0.00**	**0.920 ± 0.00**	**0.974 ± 0.00**
FedAP(+FT)	Backbone	0.873 ± 0.01	0.874 ± 0.02	0.977 ± 0.01	0.794 ± 0.02	0.832 ± 0.03	0.933 ± 0.03	0.922 ± 0.01	0.802 ± 0.03	0.951 ± 0.01
	MCDrop	0.868 ± 0.01	0.885 ± 0.01	0.980 ± 0.01	0.798 ± 0.03	0.805 ± 0.03	0.926 ± 0.02	0.927 ± 0.01	0.797 ± 0.08	0.954 ± 0.00
	Deep Ens.	0.897 ± 0.00	**0.907 ± 0.01**	0.992 ± 0.00	**0.817 ± 0.01**	0.847 ± 0.01	0.959 ± 0.00	0.935 ± 0.00	0.763 ± 0.07	0.964 ± 0.00
	FedEE	**0.921 ± 0.01**	0.906 ± 0.01	**0.995 ± 0.00**	0.783 ± 0.00	**0.870 ± 0.00**	**0.960 ± 0.00**	**0.939 ± 0.00**	**0.906 ± 0.01**	**0.968 ± 0.01**

Table 11: Performance of FedEE without personalization for classification, misdiagnosis detection, and selective prediction on the PAMAP2 dataset. Best method in bold.

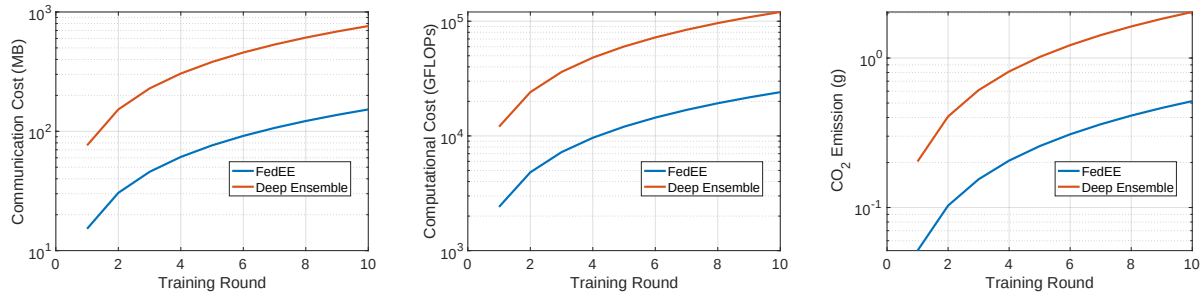
Method	PAMAP2		
	Acc. (↑)	Mis. Det. (↑)	Sel. Pred. (↑)
Backbone	0.780 ± 0.01	0.769 ± 0.02	0.866 ± 0.02
MCDrop	0.782 ± 0.02	0.817 ± 0.01	0.894 ± 0.02
Deep Ens.	0.787 ± 0.01	0.788 ± 0.01	0.877 ± 0.01
FedEE	**0.840 ± 0.01**	**0.856 ± 0.01**	**0.942 ± 0.01**
Global EE	0.789 ± 0.01	0.782 ± 0.01	0.888 ± 0.01

where e_r accounts for energy consumption of the router and e_{idle} accounts for the hardware idle energy consumption. For e_r we use the median power of router obtained from all data submitted to during 2023 to The Power Consumption Database[1], for uploading and downloading speed U and D we refer to reported values on Speedtest[2].

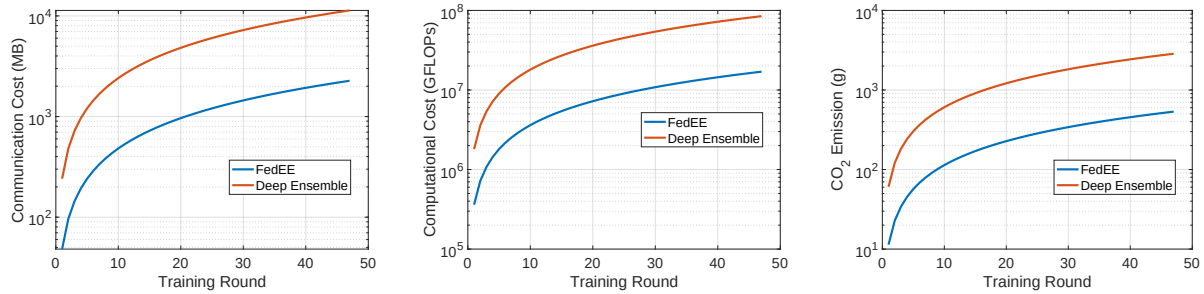
the total amount of CO2e emitted for FL is:

$$E = c_{rate}[T(e, K, R) + C(e, K, R)] \tag{11}$$

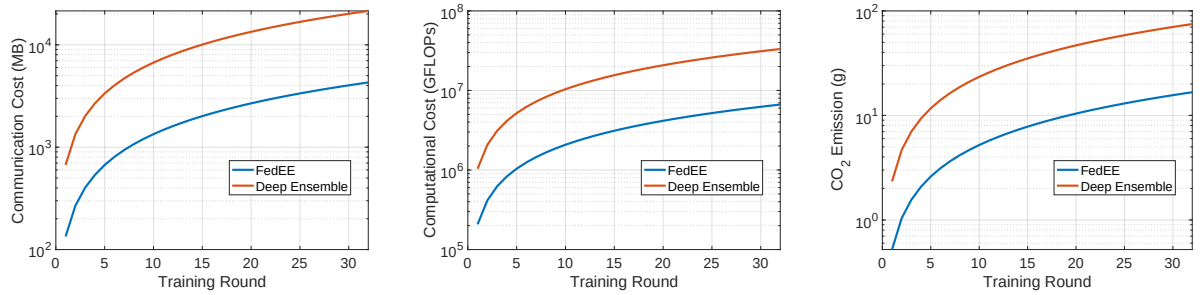
1. http://www.tpcdb.com/list.php?page=1&type=11
2. https://www.speedtest.net/global-index



(a) PAMAP2



(b) ISIC-2019



(c) PhysioNet-2016

Figure 10: Comparison of communication cost, computation cost and carbon emission of FedEE and deep ensembles.

The conversion rate from energy to carbon emission c_{rate} is estimated from official governmental websites and reports[3], also following Qiu et al. (2023).

3. https://www.climate-transparency.org/

Table 12: Estimation of energy consumption and carbon emission of FL training in three datasets, using FedAvg strategy.

		Power (W)	Clock Time(s)	Training Energy per round (Wh)	Communication Energy per round (Wh)	Total Energy (Wh)	CO_2e(g)
PAMAP2 (3-layer CNN)	Backbone	96.51	5.36	0.14	0.00	1.45	0.41
	Deep. Ens.	482.53	26.78	0.72	0.01	7.23	2.03
	FedEE	88.36	7.42	0.18	0.00	1.83	0.51
ISIC2019 (EfficientNet-b0)	Backbone	134.14	1159.53	43.21	0.00	2030.90	570.68
	Deep. Ens.	670.71	5797.66	216.03	0.02	10154.50	2853.41
	FedEE	133.00	1095.32	40.47	0.01	1902.16	534.51
PhysioNet-2016 (ResNet18)	Backbone	156.03	38.15	1.65	0.01	53.30	14.98
	Deep. Ens.	780.13	190.74	8.27	0.06	266.48	74.88
	FedEE	145.22	45.64	1.84	0.01	59.32	16.67

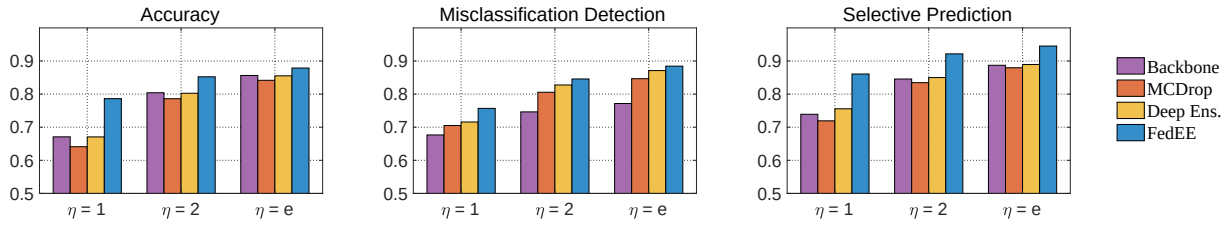


Figure 11: Performance when considering different fairness levels on PhysioNet-2016 dataset.

Averaging Client's Performance. To compare overall performance, metrics are often averaged across clients either uniformly or weighted by local data size, each with its own limitations. To address this, we propose a general averaging method defined as:

$$\lambda_\eta = \sum_{k=1}^{C} \frac{\eta^{\ln(n_k)}}{\sum_{k=1}^{C} \eta^{\ln(n_k)}} \lambda_k, \tag{12}$$

where λ_k represents the local metric and $\eta \in [1, e]$ adjusts the influence of dataset size. With $\eta = e$, the average is weighted by data size; with $\eta = 1$, it becomes a simple average. Intermediate values balance these effects. We use a weighted average ($\eta = e$) in our main experiments but also analyze different η values for a more thorough evaluation.

We conduct a detailed investigation using the PhysioNet-2016 dataset characterized by highly uneven data distribution, where one client holds over half of the data. Table 13 presents the macro average of three metrics on the PhysioNet-2016 dataset. Figure 11 shows the three metrics considering different fairness levels during result averaging. Notably, We found that FedEE exhibits substantial improvement especially with a smaller η, outperforming the baselines even more prominently. This highlights its ability to enhance performance while adhering to fairness considerations, a trend persisting when combined with other strategies.

Table 13: Performance of uncertainty estimation methods on the PhysioNet-2016 dataset, reporting predictive accuracy (Acc.), AUROC for misclassification detection (Mis. Det.) and predictive accuracy for selective prediction (Sel. Pred.), reporting macro average across clients ($\eta = 1$).

Method	Accuracy	Mis. Det.	Sel. Pred.
FedAvg			
-Backbone	0.671 ± 0.037	0.677 ± 0.038	0.739 ± 0.038
-MCDrop	0.641 ± 0.022	0.705 ± 0.035	0.719 ± 0.034
-Deep Ens.	0.671 ± 0.010	0.716 ± 0.023	0.756 ± 0.017
-FedEE	**0.786 ± 0.020**	**0.757 ± 0.021**	**0.861 ± 0.014**
FedProx			
-Backbone	0.655 ± 0.072	0.596 ± 0.033	0.699 ± 0.084
-MCDrop	0.652 ± 0.058	0.599 ± 0.049	0.696 ± 0.084
-Deep Ens.	0.727 ± 0.012	0.585 ± 0.025	0.753 ± 0.029
-FedEE	**0.785 ± 0.036**	**0.742 ± 0.028**	**0.855 ± 0.025**
FedAvg+FT			
-Backbone	0.818 ± 0.007	0.694 ± 0.032	0.862 ± 0.009
-MCDrop	0.820 ± 0.009	0.693 ± 0.081	0.864 ± 0.018
-Deep Ens.	**0.838 ± 0.012**	0.735 ± 0.030	0.892 ± 0.006
-FedEE	0.830 ± 0.026	**0.801 ± 0.019**	**0.898 ± 0.028**
FedProx+FT			
-Backbone	0.802 ± 0.012	0.720 ± 0.033	0.859 ± 0.014
-MCDrop	0.816 ± 0.018	0.727 ± 0.065	0.867 ± 0.016
-Deep Ens.	0.833 ± 0.003	0.729 ± 0.063	0.876 ± 0.008
-FedEE	**0.842 ± 0.021**	**0.761 ± 0.069**	**0.908 ± 0.021**
FedBN (+FT)			
-Backbone	0.807 ± 0.010	0.696 ± 0.025	0.856 ± 0.006
-MCDrop	0.814 ± 0.008	0.707 ± 0.068	0.865 ± 0.009
-Deep Ens.	0.835 ± 0.006	0.764 ± 0.018	0.886 ± 0.010
-FedEE	**0.881 ± 0.008**	**0.805 ± 0.011**	**0.933 ± 0.006**
FedAP (+FT)			
-Backbone	0.803 ± 0.029	0.716 ± 0.031	0.859 ± 0.029
-MCDrop	0.823 ± 0.017	0.714 ± 0.065	0.869 ± 0.021
-Deep Ens.	0.845 ± 0.009	0.682 ± 0.093	0.894 ± 0.013
-FedEE	**0.846 ± 0.013**	**0.731 ± 0.078**	**0.910 ± 0.020**