A.1 METRIC DEFINITIONS

A.1.1 Gradient variance

Gradient variance is proposed by Jiang et al. (2019) and is defined as:

$$\text{Var}(\nabla\theta_i) := \frac{1}{n}\sum_{j=1}^{n}\left(\nabla\theta_i^j - \overline{\nabla\theta_i}\right)^T\left(\nabla\theta_i^j - \overline{\nabla\theta_i}\right) \tag{A.1}$$

where θ_i^j is parameter j in layer i, $\nabla\theta_i^j$ is the gradient with respect to that parameter and $\overline{\nabla\theta_i}$ is the mean gradient of all parameters in layer i.

A.1.2 Hessian eigenvalue sum

Hessian eigenvalue sum is proposed by Chaudhari et al. (2019). The Hessian at layer i, denoted H_i, is a square matrix of second-order partial derivatives with respect to parameters at layer i. Each entry in H_i is defined as:

$$(H_i)_{jk} = \frac{\partial^2 L}{\partial\theta_i^j \partial\theta_i^k}$$

where θ_i^j and θ_i^k are parameters j and k in layer i, L is the loss and ∂. is the partial derivative with respect to that parameter. The sum of the eigenvalues is then calculated as:

$$\text{Tr}(H_i) = \sum_{p=1}^{n}\lambda_i^p \tag{A.2}$$

where λ_i^p is the p^{th} eigenvalue of H_i.

A.1.3 Sample representation similarity

Sample representation similarity is calculated via Centered Kernel Alignment (CKA) which is proposed by Kornblith et al. (2019). Its calculated as:

$$\frac{||Y_i^T X_i||_F^2}{||X_i^T X_i||_F^2 ||Y_i^T Y_i||_F^2} \tag{A.3}$$

where X_i and Y_i are sample representations after layer i coming from two distinct models.

A.1.4 F1 score

Multi-class macro-averaged F1 score is calculated as

$$\text{F1}_{\text{macro}} = \frac{1}{N}\sum_{i=1}^{N} 2 \cdot \frac{\text{precision}_i \cdot \text{recall}_i}{\text{precision}_i + \text{recall}_i} \tag{A.4}$$

where N is the number of classes, and precision_i and recall_i are the precision and recall for class i. This equation gives equal weight to all classes which is useful in imbalanced datasets like ours.

A.1.5 Algorithm Fairness

Fairness is calculated as described in (Divi et al., 2021) - the variance in individual client performances on their local test set:

$$\text{Fairness} = \frac{1}{C}\sum_{c=1}^{C}\left(P_c - \overline{P_c}\right)^2 \tag{A.5}$$

where C is the number of clients, P_c is the performance of client c and $\overline{P_c}$ is the average performance of all clients for a given personalized algorithm.

A.1.6 Algorithm Incentivization

Incentivization is defined as in (Cho et al., 2022) - the percentage of clients that outperform their local site model or global model from FedAvg:

$$\text{Incentivization} = \frac{1}{C}\sum_{c=1}^{C}\mathbb{I}\{P_c > \max(S_c, G_c)\} \tag{A.6}$$

where C is the number of clients, P_c is performance of the personalized model in client c, S_c is the performance of the local site model in client c, and G_c is the performance of the global Fed Avg model in client c

A.2 DATASETS

Table A.1 provides a description of the datasets which include: 4 image, 1 tabular and 2 NLP tasks. For FashionMNIST, EMNIST, and CIFAR-10 we create non-IID datasets by via label skew using a a Dirichlet distribution ($\alpha = 0.5$). The remaining datasets have a natural partition that we adhere to. For Sent-140 we select the top 15 users by tweet count.

Table A.1. Dataset descriptions

Dataset	Description	Classes	Partition
FashionMNIST	Fashion images	10	Label skew $Dir \sim (0.5)$
EMNIST	Handwritten digit and character images	62	Label skew $Dir \sim (0.5)$
CIFAR-10	Color images	10	Label skew $Dir \sim (0.5)$
ISIC-2019	Skin lesion images taken via dermoscopy	4	Data collected from 4 hospitals
Fed-Heart-Disease	Tabular data of patients with heart disease	5	Data collected from 5 hospitals
Sent-140	Sentiment classification of tweets	2	Data collected from 15 users
MIMIC-III	Mortality prediction using patient admission note	2	Data grouped by patient admitting diagnosis

A.3 THRESHOLD SENSITIVITY AND ROBUSTNESS

PLayer-FL uses a single hyperparameter, the threshold t, to determine the split point p from the federation-sensitivity profile (Eq. 2). Two concerns arise naturally: (i) whether the selected split is sensitive to the choice of t, and (ii) whether the aggregation of per-client sensitivity scores is robust to anomalous or noisy clients. We address both below.

A.3.1 Aggregation Strategy: Mean vs. Median

Algorithm 3 aggregates per-client federation sensitivities before determining the split point. The default strategy sums (equivalently, averages) the sensitivity scores across clients. However, a sum-based aggregation can be skewed by a single anomalous client whose sensitivity profile deviates from the majority: for example, due to a corrupted dataset, mislabeled samples, or an unusually small local dataset.

We therefore evaluate a robust alternative: replacing the sum aggregation with a *median* aggregation. Concretely, for each layer l, the server computes $\tilde{\mathcal{F}}_l(\Theta) = \text{median}_{c\in C}\big(\mathcal{F}_l(\Theta_c)\big)$ instead of $\mathcal{F}_l(\Theta) = \sum_{c\in C}\mathcal{F}_l(\Theta_c)$. The threshold comparison then proceeds identically on $\tilde{\mathcal{F}}$. This modification affects only the one-time split estimation step after epoch 1; the training loop, local updates, and server aggregation of model parameters remain unchanged.

Figure A.1 compares the two strategies across all thresholds and datasets. At the default t=2, Mean and Median agree on the split layer for *all seven datasets*. Differences between the two strategies appear only at extreme threshold values (typically $t < 1.2$), where both become overly conservative. This agreement indicates that, for the benchmarks studied, no single client dominates the aggregated sensitivity profile enough to shift the split point. Nonetheless, the median variant provides a simple safeguard for deployment in settings where client heterogeneity or data quality cannot be fully controlled.

A.3.2 Split Layer vs. Threshold

Figure A.1 shows the split layer (mode across 10 independent runs) as a function of the threshold t for all seven datasets, under both Mean (Sum) and Median aggregation.

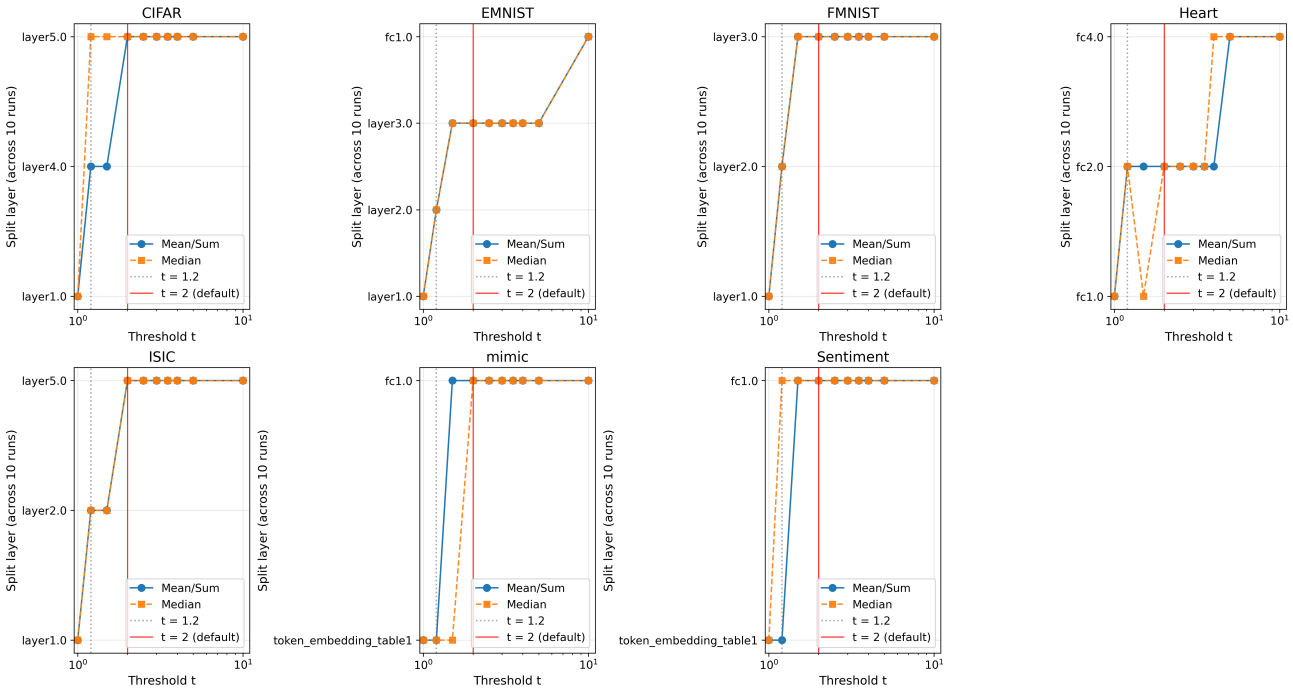


Figure A.1. Split layer (mode across 10 runs) vs. threshold t. The red solid line marks the default t=2; the grey dotted line marks t=1.2. Blue circles: Mean/Sum aggregation; orange squares: Median aggregation. For most datasets, the split layer is stable across a wide range of thresholds under both aggregation strategies.

Several observations emerge:

- **Broad stability.** For five of the seven datasets (CIFAR-10, FMNIST, ISIC, MIMIC-III, Sentiment), the split layer selected at t=2 remains unchanged across thresholds spanning roughly t=1.5 to t=10 under both aggregation strategies. This reflects the sharp spike in federation sensitivity at the transition point (Figure 4): when the jump between consecutive layers is large, even substantial changes in t do not move the split.

- **EMNIST shows a step.** For EMNIST, the Mean aggregation selects layer3 at the default t=2, but shifts to fc1 at higher thresholds (around $t \geq 5$). The Median aggregation follows a similar pattern. Both strategies agree at t=2, and the step reflects the fact that the sensitivity increase from layer3 to fc1 is moderate rather than extreme: a higher threshold naturally absorbs this increase and pushes the split one layer deeper.

- **Low-threshold regime.** At very small t (below $\approx$1.2), the split drops to early layers for some datasets (e.g., CIFAR-10 moves from layer5 to layer4 or layer1; ISIC moves from layer5 to layer2). A low threshold triggers the split at the first minor increase in sensitivity, even if that increase does not correspond to the main transition. The default t=2 avoids this regime.

- **Heart is more variable.** The Heart dataset (MLP with 4 FC layers and 2,095 parameters) shows more variability across thresholds, particularly under Median aggregation. This is consistent with the smaller model having fewer layers and less separation between consecutive sensitivity values, making the threshold comparison more sensitive to small fluctuations.

A.3.3 Split Stability Across Seeds

Table A.2 reports the standard deviation of the selected split point (expressed as a numerical layer index) across 10 independent runs at the default threshold t=2, for both aggregation strategies. A standard deviation of zero indicates that every run selected the same layer.

Table A.2. Split-point stability at default threshold t=2 across 10 independent runs. The split layer column reports the most common split layer. Std Dev is computed over the numerical layer index under Mean aggregation.

Dataset	Architecture	Split layer	Std Dev (Mean)	Std Dev (Median)
EMNIST	CNN	layer3	0.00	0.00
ISIC	CNN	layer5	0.09	0.09
CIFAR	CNN	layer5	0.14	0.12
mimic	Transformer	fc1	0.16	0.16
FMNIST	CNN	layer3	0.18	0.18
Sentiment	Transformer	fc1	0.20	0.21
Heart	MLP	fc2	0.25	0.00

The standard deviations are small across all datasets (all below 0.25), confirming that the split point is highly reproducible. For EMNIST, the split is perfectly consistent across all 10 runs. The largest variability under Mean aggregation occurs for Heart (0.25), which has only 4 layers. In this regime, minor fluctuations in gradient estimates can occasionally shift the split by one layer. Notably, the Median aggregation achieves zero standard deviation for Heart, suggesting it provides additional robustness for small models where individual client scores have more influence.

More broadly, the close agreement between Mean and Median standard deviations across all datasets reinforces the finding from Section A.3 that no single client dominates the aggregation. When outlier clients are not present, both strategies yield the same split with the same stability. The median variant thus serves as a no-cost safeguard: it preserves the default behavior under normal conditions while offering protection against anomalous clients at no additional computational or communication expense.

Implications for deployment. These results demonstrate that PLayer-FL's split-point selection is robust to (i) the choice of threshold across a wide range of values, (ii) the aggregation strategy (Mean vs. Median), and (iii) randomness across independent training runs. Since PLayer-FL's performance depends on which layers are federated, and the split is stable, the performance is correspondingly stable: consistent with the low variance observed in Table 2. Practitioners can use the default t=2 with Mean aggregation without dataset-specific tuning, and optionally switch to Median aggregation in settings where client reliability cannot be guaranteed.

A.4 Model Training

Table A.3 presents the learning rates utilized for each algorithm and Table A.4 presents the learning rate grid explored, in addition to the loss function, the number of training epochs, and the count of independent training runs. With the exception of pFedME, the AdamW optimizer was used for all algorithms. For pFedME, we adopted the specific optimizer presented by the original authors, which integrates Moreau envelopes into the training process (T Dinh et al., 2020). To account for this, we multiply the learning rate grid tested by 100 as early testing showed the pFedME optimizer demonstrated improved performance with higher learning rates. All experiments were run on 1 Tesla V100 16GB node.

A.5 Comparing models trained on non-IID data

A.5.1 Gradient variance

Figures A.2, A.3, and A.4 display the gradient variance across all datasets, corresponding to the models trained for a single epoch, final model after independent training, and final model after FL training, respectively.

Table A.3. Learning rates used

Algorithm	FMNIST	EMNIST	CIFAR	ISIC	Heart	Sent-140	MIMIC-III
Local client	$1 \cdot 10^{-3}$	$1 \cdot 10^{-3}$	$1 \cdot 10^{-3}$	$1 \cdot 10^{-3}$	$5 \cdot 10^{-2}$	$5 \cdot 10^{-4}$	$8 \cdot 10^{-5}$
FedAvg	$5 \cdot 10^{-4}$	$1 \cdot 10^{-3}$	$1 \cdot 10^{-3}$	$1 \cdot 10^{-3}$	$1 \cdot 10^{-1}$	$1 \cdot 10^{-3}$	$5 \cdot 10^{-4}$
FedProx	$5 \cdot 10^{-4}$	$5 \cdot 10^{-3}$	$1 \cdot 10^{-3}$	$1 \cdot 10^{-3}$	$5 \cdot 10^{-2}$	$1 \cdot 10^{-3}$	$5 \cdot 10^{-4}$
pFedMe	$5 \cdot 10^{-2}$	$5 \cdot 10^{-2}$	$5 \cdot 10^{-2}$	$5 \cdot 10^{-3}$	$1 \cdot 10^{-1}$	$1 \cdot 10^{-2}$	$1 \cdot 10^{-3}$
Ditto	$5 \cdot 10^{-4}$	$1 \cdot 10^{-3}$	$1 \cdot 10^{-3}$	$1 \cdot 10^{-3}$	$5 \cdot 10^{-2}$	$1 \cdot 10^{-3}$	$5 \cdot 10^{-4}$
LocalAdaptation	$5 \cdot 10^{-4}$	$1 \cdot 10^{-3}$	$1 \cdot 10^{-3}$	$1 \cdot 10^{-3}$	$1 \cdot 10^{-2}$	$1 \cdot 10^{-3}$	$5 \cdot 10^{-4}$
BABU	$1 \cdot 10^{-3}$	$1 \cdot 10^{-3}$	$5 \cdot 10^{-4}$	$1 \cdot 10^{-3}$	$1 \cdot 10^{-1}$	$8 \cdot 10^{-5}$	$5 \cdot 10^{-4}$
PLayer-FL	$1 \cdot 10^{-3}$	$1 \cdot 10^{-3}$	$1 \cdot 10^{-3}$	$5 \cdot 10^{-4}$	$1 \cdot 10^{-2}$	$8 \cdot 10^{-5}$	$3 \cdot 10^{-4}$
PLayer-FL-1	$1 \cdot 10^{-3}$	$5 \cdot 10^{-4}$	$1 \cdot 10^{-3}$	$5 \cdot 10^{-4}$	$5 \cdot 10^{-2}$	$8 \cdot 10^{-5}$	$8 \cdot 10^{-5}$
PLayer-FL+1	$1 \cdot 10^{-3}$	$1 \cdot 10^{-3}$	$5 \cdot 10^{-4}$	$1 \cdot 10^{-3}$	$5 \cdot 10^{-2}$	$1 \cdot 10^{-4}$	$5 \cdot 10^{-4}$

Table A.4. Learning rate grid, loss function, number of epochs and number of runs for each dataset

Dataset	Learning rate grid	Loss	Epochs	Runs
FashionMNIST	$1 \cdot 10^{-3}, 5 \cdot 10^{-4}, 1 \cdot 10^{-4}, 8 \cdot 10^{-5}$	Cross Entropy	75	10
EMNIST	$5 \cdot 10^{-3}, 1 \cdot 10^{-3}, 5 \cdot 10^{-4}, 1 \cdot 10^{-4}, 8 \cdot 10^{-5}$	Cross Entropy	75	10
CIFAR-10	$5 \cdot 10^{-3}, 1 \cdot 10^{-3}, 5 \cdot 10^{-4}, 1 \cdot 10^{-4}$	Cross Entropy	50	10
ISIC-2019	$1 \cdot 10^{-3}, 5 \cdot 10^{-3}, 1 \cdot 10^{-4}$	Multiclass Focal	50	3
Sent-140	$1 \cdot 10^{-3}, 5 \cdot 10^{-4}, 1 \cdot 10^{-4}, 8 \cdot 10^{-5}$	Cross Entropy	75	20
Heart	$5 \cdot 10^{-1}, 1 \cdot 10^{-1}, 5 \cdot 10^{-2}, 1 \cdot 10^{-2}, 5 \cdot 10^{-3}$	Multiclass Focal	50	50
MIMIC-III	$5 \cdot 10^{-4}, 1 \cdot 10^{-4}, 3 \cdot 10^{-4}, 8 \cdot 10^{-5}$	Multiclass Focal	50	10

A.5.2 Hessian eigenvalue sum

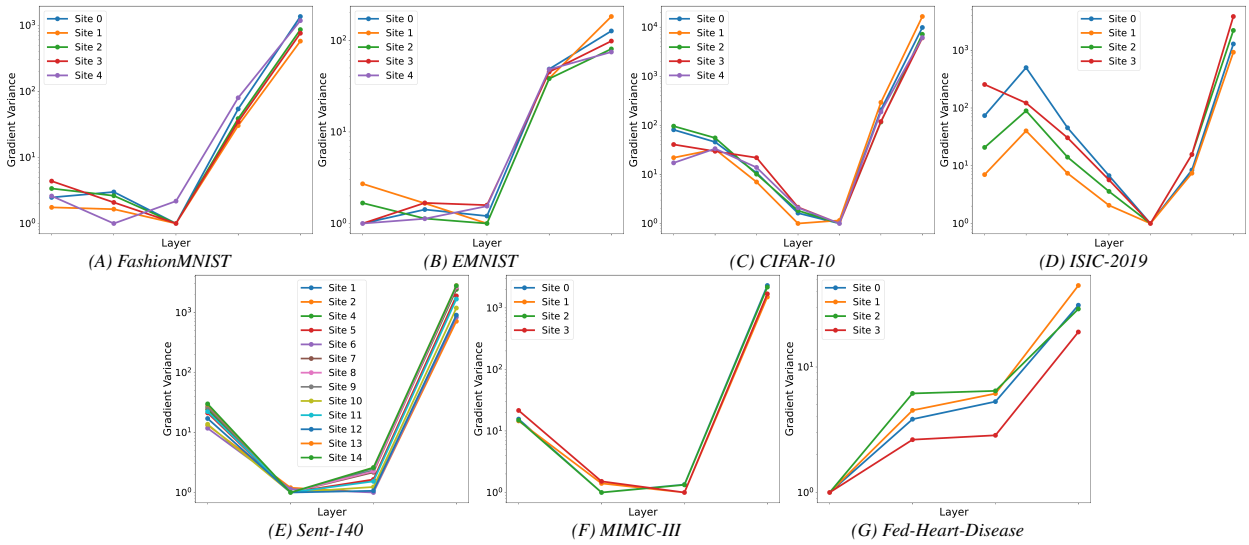
Figures A.5, A.6, and A.7 display the hessian eigenvalue sum across all datasets, corresponding to the models trained for a single epoch, final model after independent training, and final model after FL training, respectively.

A.5.3 Sample representation

Figures A.8 and A.9 display the sample representation across all datasets, corresponding to the models trained for a single epoch and the final models, respectively.

A.5.4 Federated sensitivity

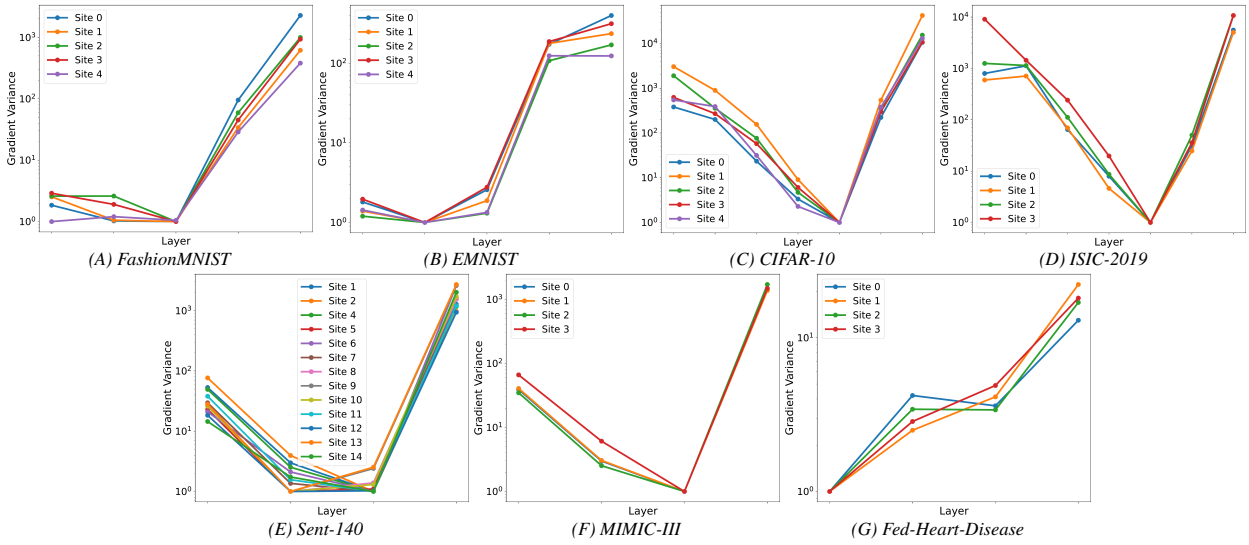
Figures A.10 and A.11 display the federated sensitivity score across all datasets for the final model after independent training, and final model after FL training, respectively.



(A) FashionMNIST *(B) EMNIST* *(C) CIFAR-10* *(D) ISIC-2019*

(E) Sent-140 *(F) MIMIC-III* *(G) Fed-Heart-Disease*

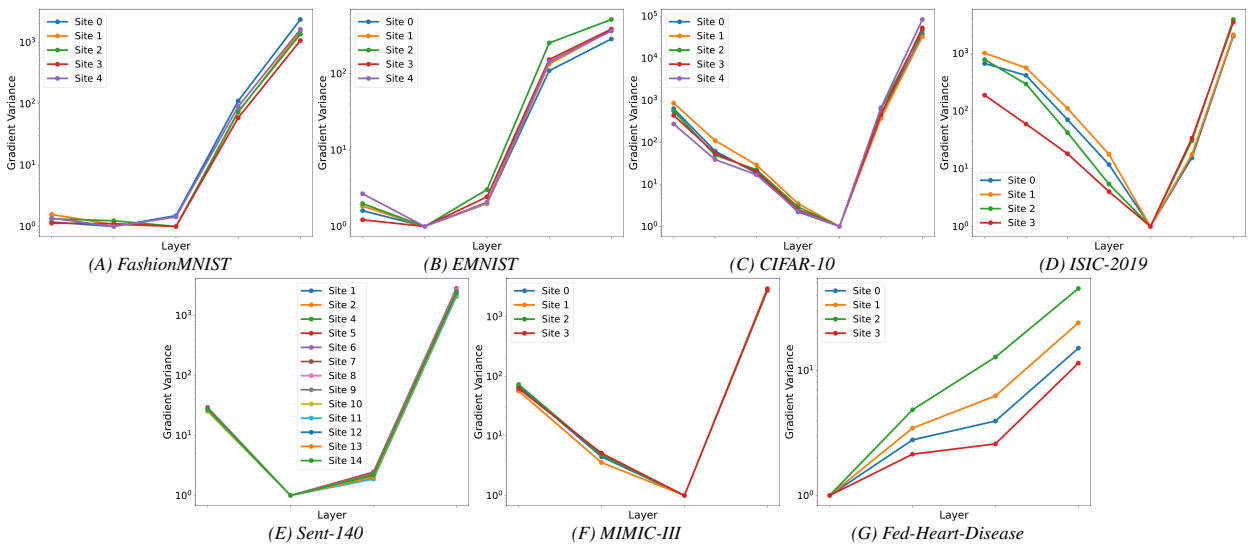
Figure A.2. **Layer gradient variance after one epoch**. All models identically initialized and independently trained on non-IID data.



(A) FashionMNIST *(B) EMNIST* *(C) CIFAR-10* *(D) ISIC-2019*

(E) Sent-140 *(F) MIMIC-III* *(G) Fed-Heart-Disease*

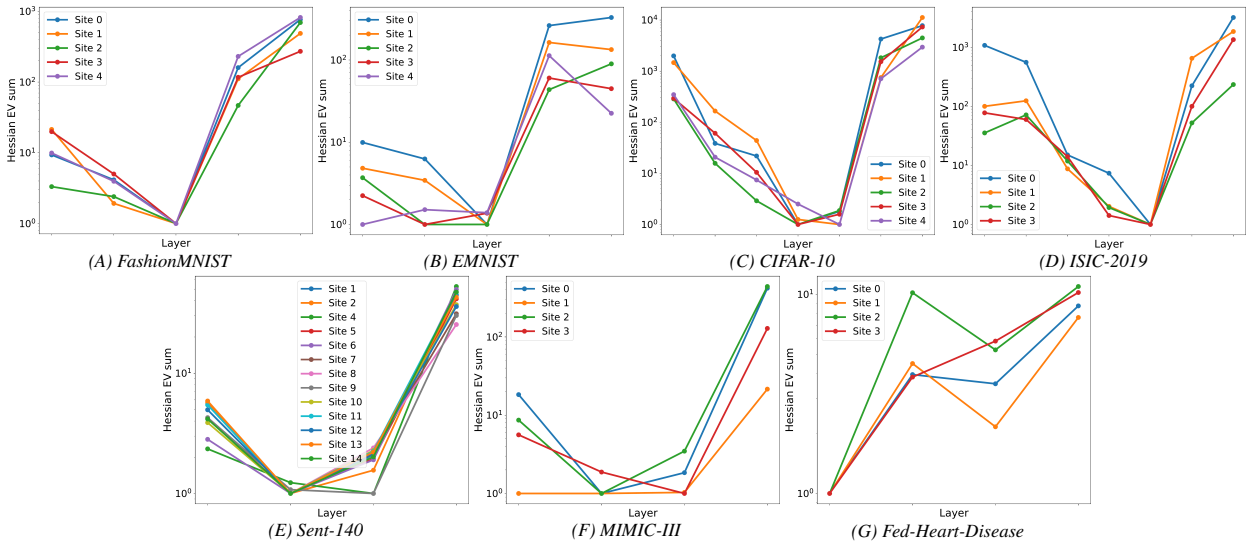
Figure A.3. **Layer gradient variance for final models**. All models identically initialized and independently trained on non-IID data.



(A) FashionMNIST *(B) EMNIST* *(C) CIFAR-10* *(D) ISIC-2019*

(E) Sent-140 *(F) MIMIC-III* *(G) Fed-Heart-Disease*

Figure A.4. **Layer gradient variance for final models**. Models trained via FL on non-IID data.



(A) FashionMNIST *(B) EMNIST* *(C) CIFAR-10* *(D) ISIC-2019*

(E) Sent-140 *(F) MIMIC-III* *(G) Fed-Heart-Disease*

Figure A.5. **Layer hessian eigenvalue sum after one epoch**. All models identically initialized and independently trained on non-IID data.

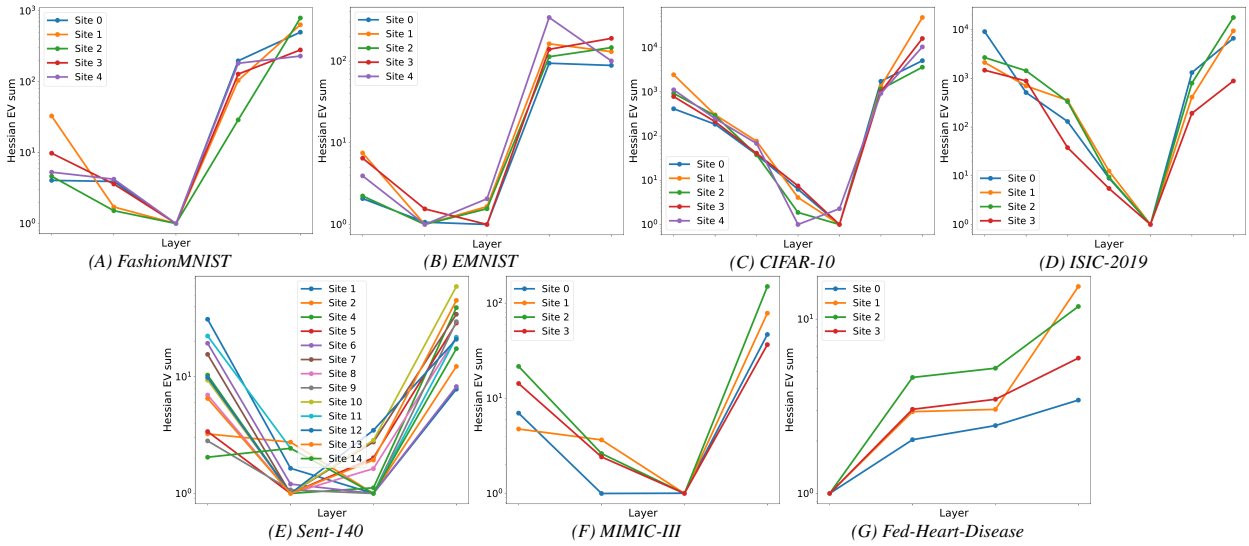


Figure A.6. **Layer hessian eigenvalue sum for the final models**. All models identically initialized and independently trained on non-IID data.

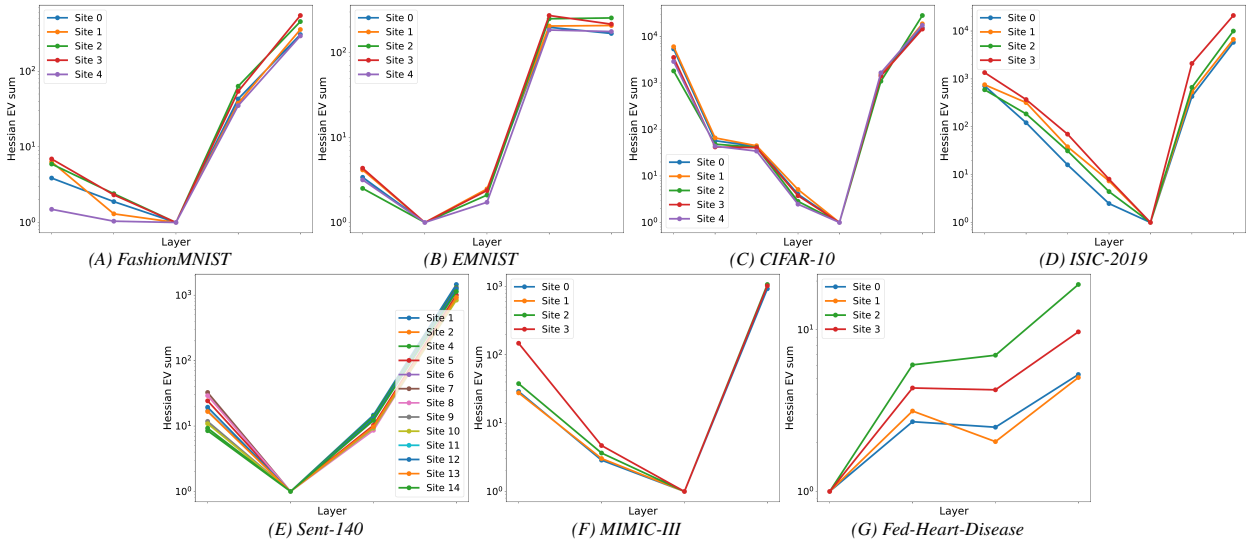


Figure A.7. **Layer hessian eigenvalue sum for the final models**. Models trained via FL on non-IID data.

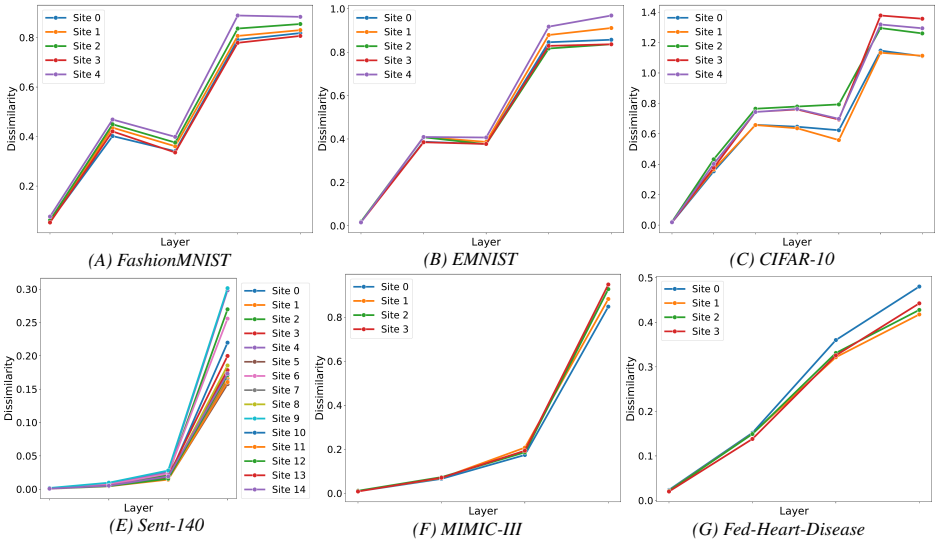


Figure A.8. **Layer sample representation similarity after one epoch**. All models identically initialized and independently trained on non-IID data.

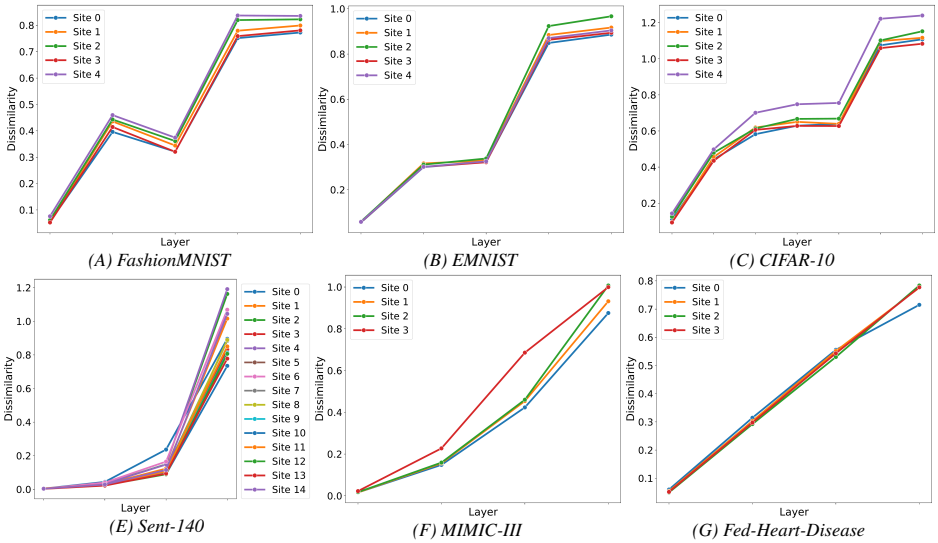


Figure A.9. **Layer sample representation similarity for final models**. All models identically initialized and independently trained on non-IID data.

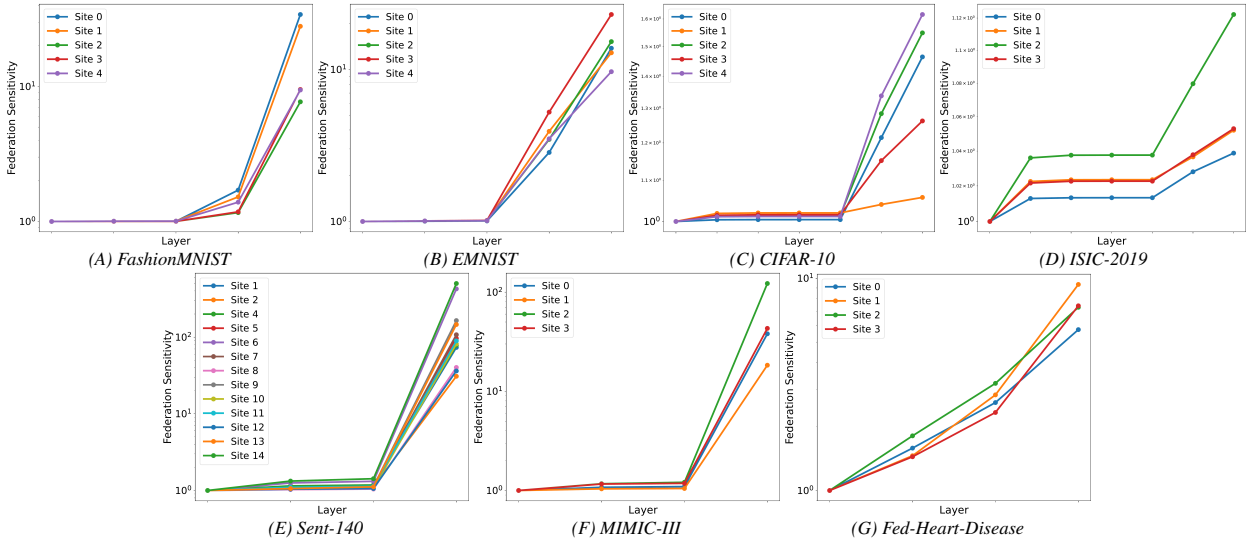


Figure A.10. **Federation sensitivity after one epoch**. All models identically initialized and independently trained on non-IID data.

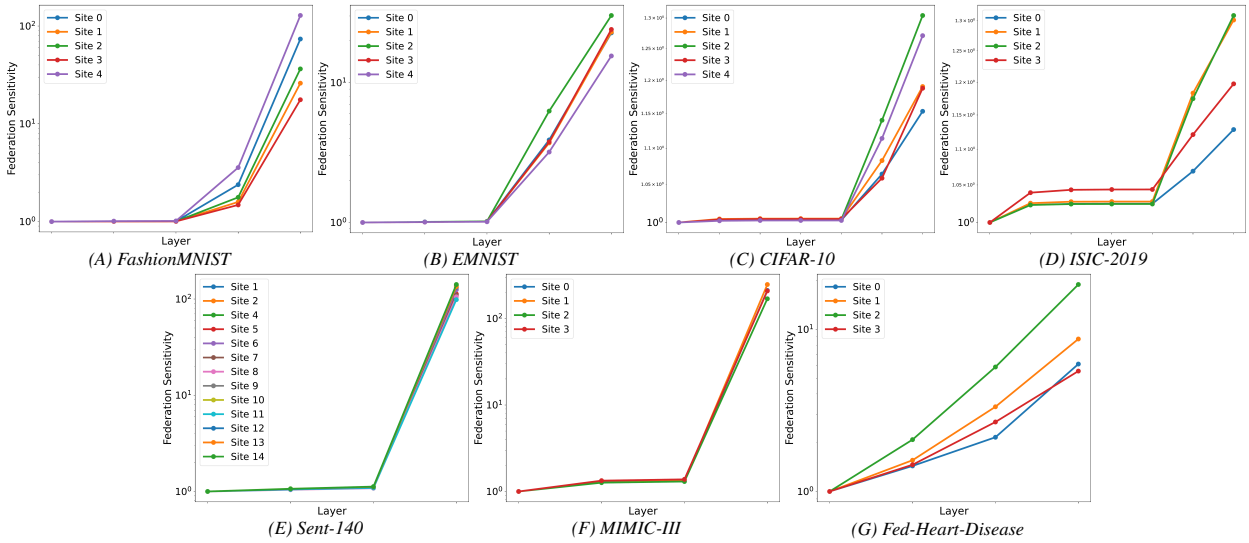


Figure A.11. **Federation sensitivity for final models**. Models trained via FL on non-IID data.

A.6 RESULTS

A.6.1 F1 score: fairness and incentive

Table A.5 shows the variance in client results and Table A.6 shows the % of clients that beat local site or FedAvg training.

Table A.5. Variance in clients' F1 score (fairness). In **bold** is fairest model. Friedman rank test p-value $< 5 \times 10^{-3}$

Algorithm	FMNIST	EMNIST	CIFAR	ISIC	Heart	Sent-140	MIMIC-III	Rank
Fedprox	$3.0 \cdot 10^{-4}$	$8.0 \cdot 10^{-4}$	$4.0 \cdot 10^{-4}$	$1.1 \cdot 10^{-2}$	$9.0 \cdot 10^{-2}$	$4.1 \cdot 10^{-3}$	$1.7 \cdot 10^{-3}$	7.7
pFedMe	$3.0 \cdot 10^{-4}$	$6.0 \cdot 10^{-4}$	$5.0 \cdot 10^{-4}$	$4.2 \cdot 10^{-3}$	$\mathbf{7.1 \cdot 10^{-2}}$	$3.7 \cdot 10^{-2}$	$\mathbf{1.4 \cdot 10^{-3}}$	4.3
Ditto	$5.0 \cdot 10^{-4}$	$6.0 \cdot 10^{-4}$	$2.0 \cdot 10^{-4}$	$3.6 \cdot 10^{-3}$	$8.2 \cdot 10^{-2}$	$3.6 \cdot 10^{-2}$	$1.6 \cdot 10^{-3}$	5.0
LocalAdaptation	$\mathbf{1.0 \cdot 10^{-4}}$	$7.0 \cdot 10^{-4}$	$4.0 \cdot 10^{-4}$	$9.0 \cdot 10^{-3}$	$8.3 \cdot 10^{-2}$	$4.1 \cdot 10^{-2}$	$2.2 \cdot 10^{-3}$	6.4
FedBABU	$2.0 \cdot 10^{-4}$	$4.0 \cdot 10^{-4}$	$4.0 \cdot 10^{-4}$	$2.1 \cdot 10^{-3}$	$8.4 \cdot 10^{-2}$	$3.5 \cdot 10^{-2}$	$2.0 \cdot 10^{-3}$	4.4
FedLP	$3.0 \cdot 10^{-4}$	$4.0 \cdot 10^{-4}$	$1.0 \cdot 10^{-3}$	$1.1 \cdot 10^{-2}$	$8.5 \cdot 10^{-2}$	$3.7 \cdot 10^{-2}$	$1.6 \cdot 10^{-3}$	6.1
FedLAMA	$2.0 \cdot 10^{-4}$	$6.0 \cdot 10^{-4}$	$5.0 \cdot 10^{-4}$	$4.7 \cdot 10^{-3}$	$8.7 \cdot 10^{-2}$	$4.1 \cdot 10^{-2}$	$2.0 \cdot 10^{-3}$	5.6
pFedLA	$4.0 \cdot 10^{-4}$	$1.0 \cdot 10^{-4}$	$\mathbf{1.0 \cdot 10^{-4}}$	$3.5 \cdot 10^{-3}$	$7.4 \cdot 10^{-2}$	$4.1 \cdot 10^{-2}$	$2.1 \cdot 10^{-3}$	5.1
PLayer-FL	$4.0 \cdot 10^{-4}$	$5.0 \cdot 10^{-4}$	$6.0 \cdot 10^{-4}$	$1.3 \cdot 10^{-3}$	$7.3 \cdot 10^{-2}$	$\mathbf{3.3 \cdot 10^{-2}}$	$1.5 \cdot 10^{-3}$	**3.8**
PLayer-FL-Random	$8.0 \cdot 10^{-4}$	$\mathbf{3.0 \cdot 10^{-4}}$	$1.0 \cdot 10^{-3}$	$\mathbf{7.0 \cdot 10^{-4}}$	$7.3 \cdot 10^{-2}$	$3.4 \cdot 10^{-2}$	$1.9 \cdot 10^{-3}$	6.5

Table A.6. Incentivized participation rate (%) using F1 score. In **bold** is model with highest IPR. Friedman rank test p-value $= 0.043$

Algorithm	FMNIST	EMNIST	CIFAR	ISIC	Heart	Sentiment	Mimic-III	Rank
FedProx	0.0	20.0	40.0	0.0	0.0	6.7	50.0	5.4
pFedMe	80.0	20.0	0.0	0.0	**50.0**	6.7	0.0	5.1
Ditto	60.0	0.0	0.0	0.0	0.0	**26.7**	25.0	5.6
LocalAdaptation	0.0	40.0	40.0	0.0	0.0	0.0	**75.0**	5.2
FedBABU	0.0	40.0	80.0	0.0	0.0	20.0	50.0	4.4
FedLP	0.0	**60.0**	0.0	0.0	0.0	6.7	50.0	6.2
FedLAMA	0.0	0.0	0.0	0.0	0.0	0.0	25.0	7.3
pFedLA	0.0	0.0	0.0	0.0	0.0	0.0	0.0	7.7
PLayer-FL	**100.0**	0.0	**100.0**	0.0	25.0	20.0	50.0	**3.4**
PLayer-FL-Random	60.0	0.0	80.0	0.0	25.0	6.7	25.0	4.8

A.6.2 Accuracy

Table A.7 shows clients' median accuracy scores (95% confidence interval), Table A.8 shows the variance in client accuracy scores and Table A.9 shows the % of clients that beat single or FedAvg training in accuracy.

A.6.3 Loss

Table A.10 shows clients' median test loss (95% confidence interval), Table A.11 shows the variance in client test loss and Table A.12 shows the % of clients that beat single or FedAvg training in test loss.

A.7 SYSTEM-LEVEL ANALYSIS

A.7.1 Communication and Computational Overhead

We analyse the communication, computational, and memory overhead of PLayer-FL relative to FedAvg.

A.7.1.1 Notation

Let L denote the number of logical layers, $|\mathcal{W}|$ the total model parameters, $|\mathcal{W}'|$ the number of non-bias parameters (used in the sensitivity metric), b the bytes per scalar (b=4 for float32), N the number of clients, R the number of communication

Table A.7. Accuracy and average rank. In **bold** is the top-performing model. Friedman rank test p-value $< 5 \times 10^{-3}$

Algorithm	FMNIST	EMNIST	CIFAR	ISIC	Heart	Sentiment	mimic	Avg Rank
Local	88.7 ± 0.9	84.2 ± 0.9	79.4 ± 1.3	**68.9 ± 1.8**	53.7 ± 1.0	79.8 ± 0.3	74.7 ± 2.1	6.7
FedAvg	89.1 ± 0.5	**86.5 ± 0.6**	80.0 ± 0.6	66.4 ± 1.2	54.1 ± 0.7	78.8 ± 0.2	74.7 ± 0.6	6.4
FedProx	89.2 ± 0.4	86.2 ± 0.9	80.2 ± 0.6	66.5 ± 1.6	52.9 ± 1.6	78.8 ± 0.2	**75.7 ± 1.7**	6.4
pFedMe	88.2 ± 0.8	85.7 ± 0.9	67.4 ± 1.1	66.8 ± 0.6	55.0 ± 1.0	80.0 ± 0.2	73.3 ± 0.9	6.9
Ditto	89.0 ± 0.6	85.6 ± 1.1	67.8 ± 0.8	66.0 ± 1.0	**55.5 ± 1.0**	80.3 ± 0.6	73.5 ± 0.9	7.0
LocalAdaptation	89.3 ± 0.8	86.3 ± 0.6	80.0 ± 1.1	66.9 ± 0.2	53.8 ± 0.8	78.9 ± 0.3	74.0 ± 2.2	6.3
FedBABU	89.2 ± 0.6	86.4 ± 0.7	80.7 ± 1.4	67.3 ± 1.6	54.2 ± 0.7	**81.0 ± 0.2**	74.9 ± 0.6	3.4
FedLP	89.1 ± 0.6	86.3 ± 0.9	78.6 ± 1.6	66.6 ± 0.7	54.1 ± 0.7	79.4 ± 0.3	74.6 ± 1.3	7.0
FedLama	86.3 ± 0.8	83.3 ± 0.6	63.4 ± 2.0	62.4 ± 1.5	54.7 ± 0.8	78.0 ± 0.1	75.2 ± 1.2	8.9
pFedLA	70.8 ± 0.2	41.2 ± 2.7	37.1 ± 2.3	56.7 ± 1.6	52.6 ± 1.4	78.0 ± 0.0	74.7 ± 2.6	11.1
PLayer-FL	**89.8 ± 0.7**	86.1 ± 0.8	**81.5 ± 1.1**	68.6 ± 0.8	55.4 ± 1.2	80.7 ± 0.7	**75.7 ± 0.6**	**2.2**
PLayer-FL-Random	89.2 ± 0.7	84.1 ± 0.8	81.2 ± 1.2	68.6 ± 0.8	54.4 ± 1.2	79.0 ± 0.4	74.5 ± 2.0	5.8

Table A.8. Variance in clients' accuracy (fairness). In **bold** is the fairest model. Friedman rank test p-value $< 5 \times 10^{-3}$

Algorithm	FMNIST	EMNIST	CIFAR	ISIC	Heart	Sentiment	mimic	Avg Rank
FedProx	$1.0 \cdot 10^{-4}$	$3.0 \cdot 10^{-4}$	$1.0 \cdot 10^{-4}$	$3.5 \cdot 10^{-2}$	$6.7 \cdot 10^{-2}$	$2.4 \cdot 10^{-2}$	$1.1 \cdot 10^{-2}$	5.7
pFedMe	$\mathbf{1.0 \cdot 10^{-5}}$	$3.0 \cdot 10^{-4}$	$4.0 \cdot 10^{-4}$	$3.3 \cdot 10^{-2}$	$4.6 \cdot 10^{-2}$	$2.1 \cdot 10^{-2}$	$1.4 \cdot 10^{-2}$	5.0
ditto	$1.0 \cdot 10^{-4}$	$2.0 \cdot 10^{-4}$	$3.0 \cdot 10^{-4}$	$3.5 \cdot 10^{-2}$	$5.7 \cdot 10^{-2}$	$2.3 \cdot 10^{-2}$	$1.3 \cdot 10^{-2}$	5.3
LocalAdaptation	$1.0 \cdot 10^{-4}$	$2.0 \cdot 10^{-4}$	$2.0 \cdot 10^{-4}$	$3.2 \cdot 10^{-2}$	$5.7 \cdot 10^{-2}$	$2.4 \cdot 10^{-2}$	$1.4 \cdot 10^{-2}$	5.5
FedBABU	$1.0 \cdot 10^{-4}$	$2.0 \cdot 10^{-4}$	$1.0 \cdot 10^{-4}$	$3.3 \cdot 10^{-2}$	$6.0 \cdot 10^{-2}$	$\mathbf{1.8 \cdot 10^{-2}}$	$1.2 \cdot 10^{-2}$	4.9
FedLP	$1.0 \cdot 10^{-4}$	$2.0 \cdot 10^{-4}$	$1.0 \cdot 10^{-4}$	$3.6 \cdot 10^{-2}$	$5.7 \cdot 10^{-2}$	$2.2 \cdot 10^{-2}$	$1.5 \cdot 10^{-2}$	5.3
FedLama	$1.0 \cdot 10^{-4}$	$2.0 \cdot 10^{-4}$	$3.0 \cdot 10^{-4}$	$4.2 \cdot 10^{-3}$	$6.3 \cdot 10^{-2}$	$2.6 \cdot 10^{-2}$	$1.3 \cdot 10^{-2}$	7.1
pFedLA	$9.0 \cdot 10^{-4}$	$6.0 \cdot 10^{-4}$	$2.3 \cdot 10^{-3}$	$6.3 \cdot 10^{-2}$	$5.1 \cdot 10^{-2}$	$2.6 \cdot 10^{-2}$	$2.5 \cdot 10^{-2}$	8.9
PLayer-FL	$1.0 \cdot 10^{-4}$	$\mathbf{1.0 \cdot 10^{-4}}$	$1.0 \cdot 10^{-4}$	$\mathbf{2.9 \cdot 10^{-2}}$	$4.7 \cdot 10^{-2}$	$1.9 \cdot 10^{-2}$	$1.3 \cdot 10^{-2}$	**2.5**
PLayer-FL-Random	$1.0 \cdot 10^{-4}$	$3.0 \cdot 10^{-4}$	$\mathbf{1.0 \cdot 10^{-4}}$	$\mathbf{2.9 \cdot 10^{-2}}$	$5.7 \cdot 10^{-2}$	$2.5 \cdot 10^{-2}$	$\mathbf{9.7 \cdot 10^{-3}}$	4.9

rounds, E the local epochs per round, D the number of local training samples, and B the batch size. PLayer-FL partitions the model at split index p (determined by federation sensitivity at threshold t): layers $1 \dots p$ are federated and layers $p{+}1 \dots L$ remain local. We write $|\mathcal{W}_F| = \sum_{l=1}^{p} |w_l|$ for the federated parameters, $|\mathcal{W}_P| = |\mathcal{W}| - |\mathcal{W}_F|$ for the personal parameters, and $\rho = |\mathcal{W}_F|/|\mathcal{W}|$ for the federated fraction.

A.7.1.2 Per-Round Communication Cost

In each round every client uploads its federated parameters and downloads the aggregated result. The per-round communication volumes are:

$$C_{\text{FA}}^{(r)} = 2\,N\,|\mathcal{W}|\,b \qquad \text{(FedAvg)}, \tag{A.7}$$

$$C_{\text{PL}}^{(r)} = 2\,N\,|\mathcal{W}_F|\,b \;=\; 2\,N\,\rho\,|\mathcal{W}|\,b \qquad \text{(PLayer-FL)}, \tag{A.8}$$

giving a per-round saving of $\Delta C^{(r)} = 2N(1{-}\rho)|\mathcal{W}|b$.

A.7.1.3 One-Time Sensitivity Overhead

PLayer-FL introduces a one-time overhead after the first epoch to compute federation sensitivity (Algorithms 2–3). This overhead has both a compute and a communication component, which we detail below.

Compute overhead. The sensitivity computation consists of three steps:

Table A.9. Incentivized Participation Rate using accuracy (%). In **bold** is model with highest IPR. Friedman rank test p-value $< 5 \times 10^{-3}$

Algorithm	FMNIST	EMNIST	CIFAR	ISIC	Heart	Sent-140	Mimic-III	Avg Rank
FedProx	20.0	**40.0**	40.0	0.0	0.0	0.0	50.0	5.6
pFedMe	40.0	**40.0**	0.0	0.0	**25.0**	6.7	0.0	5.6
Ditto	60.0	0.0	0.0	0.0	0.0	6.7	0.0	6.8
LocalAdaptation	60.0	40.0	60.0	0.0	0.0	0.0	25.0	5.3
FedBABU	80.0	20.0	60.0	0.0	0.0	**20.0**	**25.0**	4.3
FedLP	60.0	20.0	20.0	0.0	0.0	6.7	**25.0**	5.4
FedLama	0.0	0.0	0.0	0.0	0.0	0.0	0.0	8.1
pFedLA	0.0	0.0	0.0	0.0	25.0	0.0	50.0	6.5
PLayer-FL	**80.0**	0.0	**100.0**	**25.0**	**25.0**	**20.0**	**75.0**	**2.4**
PLayer-FL-Random	40.0	0.0	**100.0**	**25.0**	0.0	6.7	25.0	4.9

Table A.10. Test loss ($\times 1 \cdot 10^{-4}$) and average rank. In **bold** is the top-performing model. Friedman rank test p-value $< 5 \times 10^{-3}$

Algorithm	FMNIST	EMNIST	CIFAR	ISIC	Heart	Sentiment	mimic	Avg Rank
Local client	34.7 $\pm$ 2.0	58.6 $\pm$ 3.7	60.7 $\pm$ 4.3	**111.2 $\pm$ 0.3**	29.6 $\pm$ 0.8	43.7 $\pm$ 0.3	71.4 $\pm$ 0.8	8.3
FedAvg	30.7 $\pm$ 1.1	42.2 $\pm$ 2.8	57.4 $\pm$ 3.1	121.8 $\pm$ 1.6	27.5 $\pm$ 0.5	44.1 $\pm$ 0.2	65.8 $\pm$ 0.9	6.1
FedProx	31.5 $\pm$ 1.9	41.6 $\pm$ 2.9	56.3 $\pm$ 1.3	121.6 $\pm$ 1.1	28.4 $\pm$ 0.9	43.9 $\pm$ 0.2	65.8 $\pm$ 1.1	6.3
pFedMe	30.4 $\pm$ 2.3	42.9 $\pm$ 2.7	90.9 $\pm$ 2.8	118.0 $\pm$ 1.0	26.4 $\pm$ 0.9	42.2 $\pm$ 0.6	67.4 $\pm$ 0.8	5.1
Ditto	30.8 $\pm$ 0.2	45.0 $\pm$ 3.0	89.9 $\pm$ 3.9	120.6 $\pm$ 2.5	**26.2 $\pm$ 0.4**	42.5 $\pm$ 0.6	67.3 $\pm$ 0.1	5.9
LocalAdaptation	30.5 $\pm$ 1.9	**41.1 $\pm$ 3.4**	57.0 $\pm$ 3.0	121.4 $\pm$ 1.3	28.1 $\pm$ 0.6	43.6 $\pm$ 0.1	65.7 $\pm$ 0.6	4.9
FedBABU	30.0 $\pm$ 1.9	42.2 $\pm$ 3.5	55.2 $\pm$ 2.8	120.6 $\pm$ 1.5	29.1 $\pm$ 1.0	41.8 $\pm$ 0.2	65.5 $\pm$ 0.7	3.8
FedLP	30.1 $\pm$ 1.0	43.1 $\pm$ 3.2	61.4 $\pm$ 3.6	128.7 $\pm$ 1.5	27.3 $\pm$ 0.5	42.9 $\pm$ 0.2	65.8 $\pm$ 0.6	6.7
FedLAMA	37.6 $\pm$ 2.0	53.2 $\pm$ 2.3	102.4 $\pm$ 3.0	133.0 $\pm$ 1.8	27.2 $\pm$ 0.5	45.0 $\pm$ 0.3	**65.2 $\pm$ 0.5**	8.3
pFedLA	88.6 $\pm$ 6.9	253.4 $\pm$ 11.2	176.2 $\pm$ 10.6	151.2 $\pm$ 1.8	36.1 $\pm$ 4.9	47.2 $\pm$ 0.6	75.4 $\pm$ 0.6	12.0
PLayer-FL	**27.7 $\pm$ 2.0**	45.4 $\pm$ 3.9	**52.6 $\pm$ 2.6**	113.4 $\pm$ 0.6	26.7 $\pm$ 0.6	**41.4 $\pm$ 0.4**	66.2 $\pm$ 0.7	**3.4**
PLayer-FL-Random	33.3 $\pm$ 2.3	58.2 $\pm$ 1.8	55.6 $\pm$ 3.7	113.1 $\pm$ 1.6	27.9 $\pm$ 1.0	44.6 $\pm$ 0.7	69.9 $\pm$ 1.2	7.4

1. **Gradient availability.** The federation sensitivity metric requires the model parameters and their first-order gradients after the first epoch of training. These gradients are already computed as part of the standard training process during the last backward pass of epoch 1, so no additional forward–backward pass is required. The only requirement is that the final-batch gradients are retained rather than discarded after the optimizer step.

2. **Metric computation** (Algorithm 2). For each non-bias parameter θ_p with gradient $\nabla\theta_p$, the federation sensitivity computes $(\theta_p \cdot \nabla\theta_p)^2$, sums within each layer, and normalizes by the layer size. The dominant cost is $\mathcal{O}(|\mathcal{W}'|)$ elementwise operations (multiply, square, reduce-sum), plus L divisions for layer-wise normalization.

3. **Layer split** (Algorithm 3). The server compares the relative federation sensitivity between consecutive layers against threshold t, requiring $L-1$ comparisons.

The total additional compute cost can be formulated as:

$$\text{Compute}_{\text{sens}} = \underbrace{\mathcal{O}(|\mathcal{W}'|)}_{\text{metric (Alg. 2)}} + \underbrace{\mathcal{O}(L)}_{\text{split (Alg. 3)}} . \tag{A.9}$$

This expression is approximate; the actual cost depends on implementation details such as operator fusion, memory allocation patterns, and framework-level overhead, but captures the dominant terms.

Since the gradients are reused from the final training step of epoch 1, the sensitivity computation adds only an elementwise

Table A.11. Variance in clients' test loss (fairness). In **bold** is the fairest model. Friedman rank test p-value $< 5 \times 10^{-3}$

Algorithm	FMNIST	EMNIST	CIFAR	ISIC	Heart	Sent-140	MIMIC-III	Avg Rank
FedProx	$1.1 \cdot 10^{-3}$	$1.7 \cdot 10^{-3}$	$1.1 \cdot 10^{-3}$	$3.8 \cdot 10^{-1}$	$1.8 \cdot 10^{-2}$	$5.5 \cdot 10^{-2}$	$4.7 \cdot 10^{-2}$	6.4
pFedMe	$5.0 \cdot 10^{-4}$	$2.3 \cdot 10^{-3}$	$1.8 \cdot 10^{-3}$	$3.6 \cdot 10^{-1}$	$1.3 \cdot 10^{-2}$	$5.2 \cdot 10^{-2}$	$3.5 \cdot 10^{-2}$	4.1
Ditto	$6.0 \cdot 10^{-4}$	$2.2 \cdot 10^{-3}$	$2.3 \cdot 10^{-4}$	$3.8 \cdot 10^{-1}$	$1.7 \cdot 10^{-2}$	$5.2 \cdot 10^{-2}$	$3.8 \cdot 10^{-2}$	4.6
LocalAdaptation	$1.0 \cdot 10^{-3}$	$2.1 \cdot 10^{-3}$	$1.2 \cdot 10^{-3}$	$4.0 \cdot 10^{-1}$	$1.9 \cdot 10^{-2}$	$5.5 \cdot 10^{-2}$	$4.8 \cdot 10^{-2}$	7.1
FedBABU	$4.0 \cdot 10^{-4}$	$\mathbf{1.2 \cdot 10^{-3}}$	$7.0 \cdot 10^{-4}$	$4.1 \cdot 10^{-1}$	$2.1 \cdot 10^{-2}$	$\mathbf{4.6 \cdot 10^{-2}}$	$4.5 \cdot 10^{-2}$	4.4
FedLP	$7.0 \cdot 10^{-4}$	$1.9 \cdot 10^{-3}$	$8.0 \cdot 10^{-4}$	$4.3 \cdot 10^{-1}$	$1.7 \cdot 10^{-2}$	$5.2 \cdot 10^{-2}$	$4.3 \cdot 10^{-2}$	5.1
FedLAMA	$\mathbf{2.0 \cdot 10^{-4}}$	$2.9 \cdot 10^{-3}$	$2.4 \cdot 10^{-3}$	$4.6 \cdot 10^{-1}$	$1.8 \cdot 10^{-2}$	$5.3 \cdot 10^{-2}$	$4.3 \cdot 10^{-2}$	6.7
pFedLA	$1.0 \cdot 10^{-3}$	$2.3 \cdot 10^{-2}$	$5.1 \cdot 10^{-2}$	$5.7 \cdot 10^{-1}$	$2.6 \cdot 10^{-1}$	$5.1 \cdot 10^{-2}$	$\mathbf{3.2 \cdot 10^{-2}}$	7.6
PLayer-FL	$7.0 \cdot 10^{-4}$	$2.5 \cdot 10^{-3}$	$4.0 \cdot 10^{-4}$	$\mathbf{3.2 \cdot 10^{-1}}$	$\mathbf{1.4 \cdot 10^{-2}}$	$5.1 \cdot 10^{-2}$	$4.2 \cdot 10^{-2}$	**3.6**
PLayer-FL-Random	$8.0 \cdot 10^{-4}$	$4.1 \cdot 10^{-3}$	$\mathbf{1.0 \cdot 10^{-4}}$	$\mathbf{3.2 \cdot 10^{-1}}$	$1.6 \cdot 10^{-2}$	$6.1 \cdot 10^{-2}$	$4.0 \cdot 10^{-2}$	5.2

Table A.12. Incentivized Participation Rate using test loss (%) and Average Algorithm Rank. Friedman rank test p-value $= 0.084$

Algorithm	FMNIST	EMNIST	CIFAR	ISIC	Heart	Sent-140	Mimic-III	Avg Rank
FedProx	40.0	**80.0**	40.0	0.0	25.0	33.3	25.0	5.1
pFedMe	40.0	40.0	0.0	0.0	**100.0**	33.3	0.0	5.7
Ditto	40.0	0.0	0.0	0.0	75.0	33.3	0.0	6.5
LocalAdaptation	20.0	60.0	80.0	0.0	25.0	40.0	25.0	4.8
FedBABU	80.0	40.0	60.0	0.0	0.0	40.0	25.0	4.6
FedLP	60.0	**80.0**	20.0	0.0	75.0	40.0	50.0	3.6
FedLAMA	0.0	0.0	0.0	0.0	50.0	20.0	**75.0**	6.7
pFedLA	0.0	0.0	0.0	0.0	0.0	6.7	0.0	8.5
PLayer-FL	**80.0**	0.0	**100.0**	**50.0**	**100.0**	40.0	0.0	**3.5**
PLayer-FL-Random	0.0	0.0	**100.0**	**50.0**	50.0	20.0	0.0	6.0

scan of the parameter–gradient products and a small number of per-layer operations on top of the standard training that both FedAvg and PLayer-FL perform.

Communication overhead. The sensitivity step requires each client to send its L per-layer sensitivity scores to the server for aggregation (Algorithm 3). The one-time communication cost is:

$$C_{\text{sens}}^{(\text{comm})} = N \cdot L \cdot b. \tag{A.10}$$

For concrete reference: CIFAR-10 (N=5, L=7, b=4) incurs $5 \times 7 \times 4 = 140$ bytes; MIMIC-III (N=4, L=7) incurs $4 \times 7 \times 4 = 112$ bytes. This is negligible compared to per-round volumes of tens of megabytes to gigabytes.

A.7.1.4 Total Communication Cost

Combining the per-round cost over R rounds with the one-time sensitivity communication:

$$C_{\text{FA}} = 2\,N\,|\mathcal{W}|\,b\,R, \tag{A.11}$$

$$C_{\text{PL}} = 2\,N\,|\mathcal{W}_F|\,b\,R \;+\; N \cdot L \cdot b. \tag{A.12}$$

The break-even round (where PLayer-FL's cumulative communication first falls below FedAvg's) is:

$$r^* = \frac{N \cdot L \cdot b}{2\,N\,(1-\rho)\,|\mathcal{W}|\,b} = \frac{L}{2\,(1-\rho)\,|\mathcal{W}|}. \tag{A.13}$$

Since $|\mathcal{W}| \gg L$ for any practical model, $r^* \approx 0$ for all datasets: the communication overhead of federation sensitivity is amortised immediately within the first round. For example, CIFAR-10 has $r^* = 7\,/\,(2 \times 0.0132 \times 3{,}963{,}094) \approx 6.7 \times 10^{-5}$ rounds.

A.7.1.5 Per-Round Computational Complexity

Per round, both methods perform N local training passes of E epochs. Each epoch consists of $\lceil D/B \rceil$ iterations, each involving a forward pass, a backward pass, and a parameter update. For both SGD and AdamW (used in our experiments, Table A.4), the forward and backward passes dominate at $\mathcal{O}(B \cdot |\mathcal{W}|)$ per iteration. The parameter update is $\mathcal{O}(|\mathcal{W}|)$ for both optimizers (AdamW adds a constant factor of elementwise operations for moment updates, bias correction, and weight decay, but remains $\mathcal{O}(|\mathcal{W}|)$). The cost of one local epoch is therefore:

$$F(|\mathcal{W}|) = \left\lceil \frac{D}{B} \right\rceil \cdot \mathcal{O}(B \cdot |\mathcal{W}|) = \mathcal{O}(D \cdot |\mathcal{W}|). \tag{A.14}$$

The per-round complexity, including E local epochs and the aggregation step, is:

$$\text{FedAvg:} \quad \mathcal{O}\big(N \cdot E \cdot D \cdot |\mathcal{W}| \; + \; |\mathcal{W}|\big), \tag{A.15}$$

$$\text{PLayer-FL:} \quad \mathcal{O}\big(N \cdot E \cdot D \cdot |\mathcal{W}| \; + \; |\mathcal{W}_F|\big). \tag{A.16}$$

The local training term $N \cdot E \cdot D \cdot |\mathcal{W}|$ dominates completely, making the aggregation difference ($|\mathcal{W}|$ vs $|\mathcal{W}_F|$) negligible in practice. Both methods have effectively identical per-round compute cost. The only additional cost is the one-time sensitivity computation (Eq. A.9), executed once after epoch 1.

A.7.1.6 Memory Complexity

Both FedAvg and PLayer-FL train the *full* model locally, the split only determines which parameters are communicated to the server for aggregation. Consequently, both methods maintain similar memory footprints during training. The dominant contributions are:

$$M = \underbrace{|\mathcal{W}| \cdot b}_{\text{model parameters}} + \underbrace{2 \cdot |\mathcal{W}| \cdot b}_{\text{optimizer states (AdamW)}} + \underbrace{\mathcal{O}(B \cdot |\mathcal{W}|)}_{\text{activations}} = \mathcal{O}(B \cdot |\mathcal{W}|). \tag{A.17}$$

This expression captures the leading-order terms; actual memory usage also depends on implementation details such as intermediate buffers, batch normalization statistics, and framework overhead. The optimizer states account for the first and second moment estimates (m_t and v_t) required by AdamW, which are maintained for all $|\mathcal{W}|$ parameters regardless of the federation strategy. PLayer-FL adds only:

- L scalar federation-sensitivity scores per client, and
- a single integer split index p,

for a total of $\mathcal{O}(L)$ additional memory, negligible for any practical architecture. The key point is that PLayer-FL and FedAvg share the same asymptotic memory complexity; the difference is limited to the $\mathcal{O}(L)$ overhead for storing the sensitivity scores.

Theoretical summary. PLayer-FL introduces no additional per-round computation or memory beyond standard FedAvg training. Its sole overhead is a one-time sensitivity analysis after the first epoch, consisting of $\mathcal{O}(|\mathcal{W}'|)$ operations for the metric computation and $\mathcal{O}(L)$ for the layer split, with $\mathcal{O}(L)$ additional memory. The one-time communication cost of transmitting L per-layer sensitivity scores ($N \cdot L \cdot b$ bytes total) is negligible (typically under 500 bytes) and amortised immediately ($r^* \approx 0$). Per-round communication savings are architecture-dependent: determined by the fraction of parameters that fall after the transition point identified by the federation-sensitivity metric.

A.7.1.7 Empirical Communication Volumes

We validate the theoretical analysis by instantiating the actual model architectures and measuring the size of the exchanged `state_dict` for each dataset. The split point p is determined by federation sensitivity at the default threshold t=2. Table A.13 reports per-round and total communication volumes.

The measured volumes confirm the theoretical predictions. As a concrete validation, consider MIMIC-III (N=4, R=25,

Table A.13. Communication volume: FedAvg vs PLayer-FL. Per-round values reflect upload + download ($2N|\mathcal{W}_{(\cdot)}|b$). Savings percentage equals $100(1-\rho)$.

Dataset	N	R	Split layer (p)	Per-Round		Total		Savings
				FedAvg	PLayer-FL	FedAvg	PLayer-FL	(%)
EMNIST	5	75	layer3 (3)	42.3 MB	41.0 MB	3.2 GB	3.1 GB	3.0
FMNIST	5	100	layer3 (3)	42.1 MB	41.0 MB	4.2 GB	4.1 GB	2.5
CIFAR	5	100	layer5 (5)	158.5 MB	156.4 MB	15.9 GB	15.6 GB	1.3
ISIC	4	60	layer5 (5)	54.4 MB	50.2 MB	3.3 GB	3.0 GB	7.8
Sentiment	15	50	fc1 (5)	1.7 GB	1.5 GB	86.7 GB	72.5 GB	16.3
Heart	4	20	fc2 (2)	67.0 KB	50.2 KB	1.3 MB	1.0 MB	25.1
MIMIC-III	4	25	fc1 (5)	554.7 MB	479.1 MB	13.9 GB	12.0 GB	13.6

$|\mathcal{W}|$=17,332,994, b=4, split at fc1 with $|\mathcal{W}_F|$=14,971,392):

$$
\begin{aligned}
C_{\text{FA}}^{(r)} &= 2 \times 4 \times 17{,}332{,}994 \times 4 = 554{,}655{,}808 \text{ bytes} \approx 554.7 \text{ MB}, \\
C_{\text{PL}}^{(r)} &= 2 \times 4 \times 14{,}971{,}392 \times 4 = 479{,}084{,}544 \text{ bytes} \approx 479.1 \text{ MB}, \\
\Delta C^{(r)} &= 554{,}655{,}808 - 479{,}084{,}544 = 75{,}571{,}264 \text{ bytes} \approx 75.6 \text{ MB}, \\
\text{Savings} &= 1 - \rho = 1 - 14{,}971{,}392 \,/\, 17{,}332{,}994 = 13.6\%, \\
C_{\text{sens}}^{(\text{comm})} &= 4 \times 7 \times 4 = 112 \text{ bytes}, \\
r^* &= 7 \,/\, (2 \times 0.1362 \times 17{,}332{,}994) \approx 1.5 \times 10^{-6} \text{ rounds}.
\end{aligned}
$$

The experimental measurements in Table A.13 (554.7 MB and 479.1 MB per round, 13.6% savings) match these theoretical values, confirming that the formulas in Eqs. A.7–A.13 hold.

The magnitude of savings varies across architectures. For CNN models (EMNIST, FMNIST, CIFAR-10), the federation-sensitivity metric places the transition point at or near the last convolutional layer. Because these architectures use compact fully-connected heads, the personal fraction $|\mathcal{W}_P|$ is small relative to the total model, yielding modest per-round savings (1–8%). In contrast, for Transformer-based models (Sentiment, MIMIC-III) and the MLP (Heart), a larger fraction of parameters falls after the transition point, producing savings of 13–25%.

A.7.1.8 Communication as a Function of the Split Point

Figure A.12 shows how per-round communication scales with the fraction of federated parameters across all possible split points. The relationship is linear by construction (Eq. A.8). The markers indicate the split selected by PLayer-FL at t=2 and the FedAvg baseline (all layers federated). This visualisation clarifies the trade-off space: moving the split earlier reduces communication but also reduces the set of layers benefiting from cross-client aggregation.

A.7.1.9 Measured Sensitivity Computation Overhead

Table A.14 reports the wall-clock time for the one-time federation-sensitivity metric computation (Algorithm 2) after the first epoch. Since the metric reuses gradients already computed during the last training step of epoch 1, the overhead consists solely of the elementwise parameter–gradient operations and per-layer aggregation. Measurements were taken on CPU with 10 repetitions after 5 warm-up iterations (median reported).

The metric computation completes in under 90 ms even for the largest model (MIMIC-III, 17M parameters) on CPU, and scales approximately linearly with the number of parameters. This confirms that the one-time sensitivity overhead is negligible compared to even a single epoch of training.

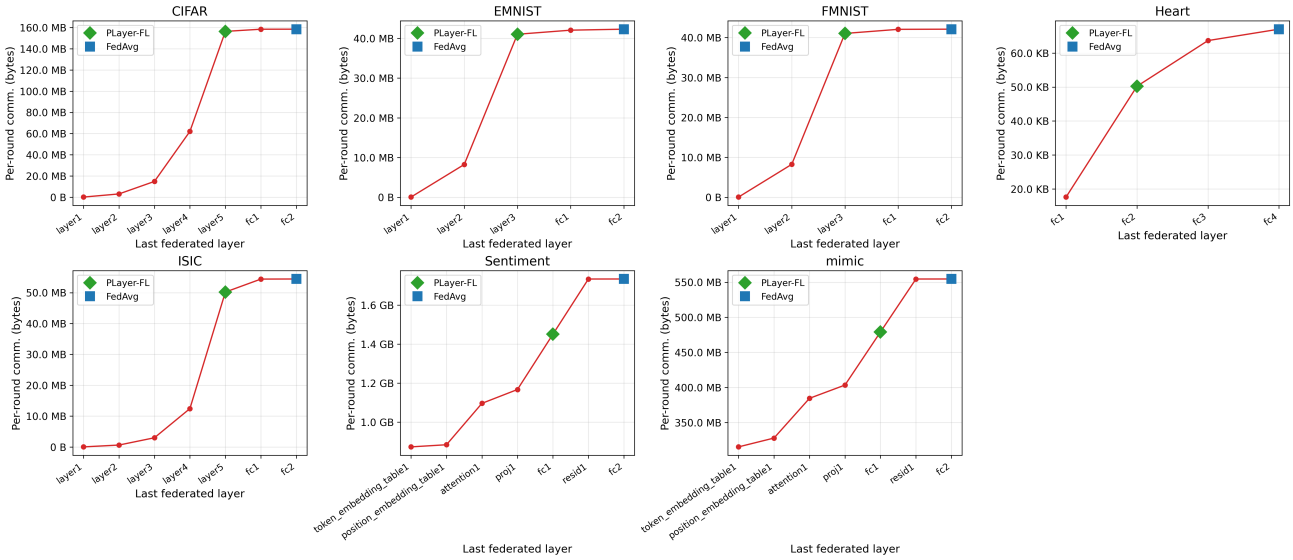


Figure A.12. Per-round communication volume as a function of the fraction of federated parameters. Each point corresponds to a possible split index p. Blue square: FedAvg ($p{=}L$, all layers federated); Green diamond: PLayer-FL (p selected by federation sensitivity at $t{=}2$).

Table A.14. One-time federation-sensitivity metric computation overhead (CPU, median of 10 runs). The overhead is incurred once after epoch 1 and reuses gradients from training.

Dataset	$\lvert\mathcal{W}\rvert$	Metric computation (ms)
EMNIST	1,058,010	4.6
FMNIST	1,052,758	5.5
CIFAR	3,963,094	16.5
ISIC	1,700,932	5.8
Sentiment	14,447,618	48.1
Heart	2,095	0.2
MIMIC-III	17,332,994	88.9