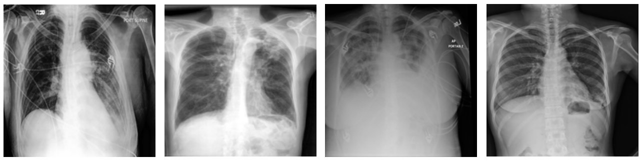
Appendix



(a) CheX (b) OpenI (c) RSNA (d) VinBigdata

Figure 1: Representative chest X-ray samples from the four domains in Chest-X-Ray-4: (a) CheXpert, (b) OpenI, (c) RSNA, and (d) VinBigdata. The visible differences in scanner type, contrast, and patient positioning illustrate the real-world heterogeneity that federated models must handle.

Chest-X-Ray-4 Dataset Details

We curate a four-domain chest radiograph suite that mirrors typical inter-hospital heterogeneity while respecting privacy constraints:

- **CheXpert (C)** – Stanford tertiary-care hospitals, Western USA (12-bit DR systems) (Irvin et al. 2019).
- **OpenI (O)** – National Library of Medicine public archive; older CR scanners and diverse capture protocols (Demner-Fushman et al. 2016).
- **RSNA (R)** – Pneumonia-screening programme across children's hospitals, NortheasternUSA; paediatric bias (Shih et al. 2019).
- **VinBigdata (V)** – Provincial hospital, Vietnam; low-resource setting with mixed analogue–digital pipeline (Nguyen et al. 2022).

All images are resized to 224×224 and re-labelled into the binary task Lung Opacity vs. Normal. The sample is shown in the Figure 1. To ensure comparable on-device workloads we down-sample each domain to 458 training images and hold out an equal-size (OpenI: 116 due to data scarcity) test split that is never seen during training. The federated setting therefore involves four clients with strictly non-overlapping distributions P_C, P_O, P_R, and P_V. Detailed class counts and preprocessing scripts are available in our public code repository.

Ethical Considerations and Responsible Data Use: The curation and use of the Chest-X-Ray-4 dataset were guided by strict ethical principles, recognizing the sensitive nature of patient data and the potential for social impact. It is crucial to note that all four source datasets are publicly available, de-identified research collections that were released under their own institutional review board (IRB) or equivalent ethical approvals. We rely on the de-identification measures performed by the original data curators, ensuring that no personally identifiable information was handled in our study. Our use of these datasets is for research purposes, adhering to the data use agreements of each source. Furthermore, our research aims to avoid data centralization through federated learning, providing additional privacy protection for future studies.

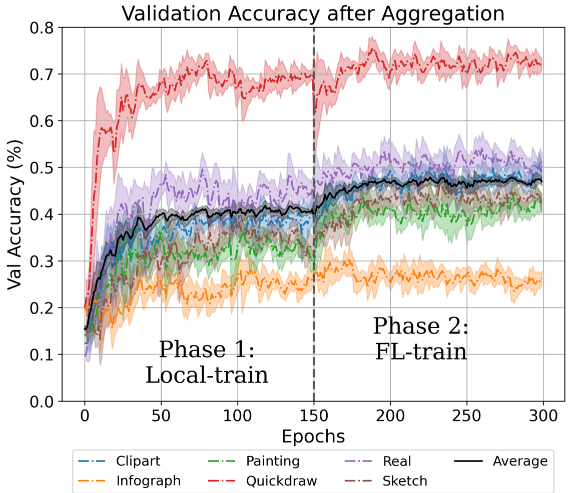


Figure 2: Per-domain validation accuracy curves on the DomainNet dataset throughout the two-phase training process. The plot illustrates the transition from Phase 1 (local-only pre-training, epochs 0-150) to Phase 2 (federated collaboration, epochs 151-300). Performance diverges significantly in Phase 1 based on domain characteristics, while Phase 2 shows knowledge sharing leading to improved performance for most domains, particularly those that initially struggled.

Detailed Training Dynamics

Figure 2 provides a detailed, per-domain visualization of our two-phase training process on the DomainNet benchmark. This plot complements the aggregated results presented in the main paper by illustrating the learning trajectory of each individual client.

In Phase 1 (epochs 0-150), each client trains its model exclusively on its local data. This results in performance trajectories that are highly dependent on the characteristics of the individual domain. For example, the *Quickdraw* domain, with its distinct and clean line-art style, allows the model to achieve high accuracy rapidly. In contrast, domains like *Infograph* and *Painting* present greater learning challenges, resulting in lower initial performance.

The transition to Phase 2 (epoch 151) marks the beginning of federated collaboration, where clients begin to train and aggregate the shared branch. The effect is immediate: the aggregated knowledge starts to benefit multiple clients. Notably, domains that were previously under-performing, such as *Sketch*, *Real*, and *Painting*, exhibit a clear and sustained increase in accuracy. This demonstrates that the shared branch is successfully capturing and distributing valuable domain-agnostic features, allowing weaker clients to learn from the knowledge of the entire federation.

This visualization empirically supports our central claim that the pFedDB framework allows clients to first build a strong local foundation and then leverage federated knowledge to enhance performance, without suffering from the instability or catastrophic forgetting seen in simpler aggregation schemes.

Table 1: Per-domain accuracy for different cut depths l_c. Left: Office-Caltech-10 with *AlexNet*; Right: Chest-X-Ray-4 with *DenseNet-121*. *Param* = MB uploaded per round; *Ratio* = communicated fraction of total parameters.

Office-Caltech-10 (AlexNet)

l_c	Param	Ratio	A	C	D	W	Avg.
5	9.4	0.04	62.0	41.8	81.3	82.1	66.8
6	153.5	0.71	62.5	50.1	90.6	90.5	**73.4**
7	217.5	0.99	62.0	50.1	87.5	88.8	72.1
8	217.7	1.00	61.5	43.6	90.6	85.4	70.3

Chest-X-Ray-4 (DenseNet-121)

l_c	Param	Ratio	C	O	R	V	Avg.
14	1.4	0.05	84.2	73.3	86.0	80.1	80.9
39	5.4	0.20	85.4	71.6	84.9	83.0	81.2
88	18.2	0.68	85.4	73.3	86.0	89.1	**83.4**
120	26.5	0.99	85.4	70.7	86.0	88.9	82.7

Detailed Per-Domain Results of the Cut-Depth Ablation

Table 1 break down the cut-depth study by listing the accuracy of every individual domain alongside the overall mean. The expanded view confirms the trend discussed in the main text: allocating roughly 70% of the total parameters to the shared branch delivers the best balance between communication cost and accuracy. Specifically, splitting AlexNet after the first fully connected layer ($l_c = 6$, 71% of parameters) and DenseNet-121 after Block 3 ($l_c = 88$, 68%) yields the highest average accuracy while keeping per-domain performance uniformly strong. Shallower cuts (e.g., $l_c = 5$ for AlexNet or $l_c = 14$ for DenseNet-121) still beat full-model FedAvg at a fraction of the bandwidth, illustrating the graceful accuracy–bandwidth trade-off enabled by the dual-branch design. Deeper cuts that transmit virtually the entire network offer no additional benefit and sometimes hurt certain domains, reinforcing the practical guideline to place the split near the 70% parameter mark.

Analysis with Standard-Deviation Reporting

As shown in Tables 2-5, we augmented all main-result tables with the standard deviation of each score (values in parentheses). The first row is the single baseline network. Two patterns emerge. (i) The superiority of pFedDB remains statistically meaningful: on every benchmark it either attains the best or tied-best mean accuracy, and its lead over the next-best method consistently exceeds or comparable to the reported standard deviations, indicating the improvements are not due to random fluctuation. (ii) Our pFedDB converges more consistently: its standard deviations are among the lowest across datasets, showing that the dual-branch design not only boosts performance but also reduces variability across random seeds and train-test splits. These additional statistics therefore strengthen the evidence for the reliability and practical utility of our approach in multi-domain settings.

Computing Infrastructure

All experiments were conducted using Vscode and pytorch with the following specifications:

- GPU: 2 * NVIDIA 3080 with 12GB Memory
- CPU: AMD EPYC 7313
- Operating System: Ubuntu 20.04 LTS
- RAM: 1TB

Table 2: Chest-X-Ray-4 multi-domain classification accuracy (%). Mean (top) and standard deviation (bottom).

Method	C	O	R	V	Avg
DenseNet121	82.3	68.1	83.6	88.9	80.7
	(1.3)	(1.1)	(0.5)	(2.6)	(0.6)
FedAvg	82.2	67.6	84.1	84.7	79.6
	(2.0)	(2.2)	(1.2)	(1.2)	(0.7)
FedProx	79.3	65.5	85.9	85.5	79.0
	(0.8)	(4.9)	(0.1)	(2.8)	(1.7)
FedBN	82.9	67.2	86.5	87.6	81.0
	(1.8)	(2.1)	(1.5)	(0.4)	(0.1)
pFedSD	81.1	72.3	**86.5**	84.5	81.1
	(1.7)	(2.7)	**(1.4)**	(2.2)	(0.7)
FedSSD	84.2	72.4	83.2	84.9	81.2
	(0.8)	(1.0)	(0.8)	(0.5)	(0.4)
FedWon	84.5	72.3	85.2	83.8	81.5
	(3.4)	(3.7)	(1.0)	(1.9)	(0.9)
pFedDB	**85.4**	**73.3**	86.0	**89.1**	**83.4**
	(2.5)	**(1.8)**	(1.0)	**(1.6)**	**(0.7)**

Table 3: Digits-Five multi-domain classification accuracy (%). Mean (top) and standard deviation (bottom).

Method	Mn	Sv	Up	Sy	MM	Avg
Simple-CNN	94.4	65.3	95.2	80.3	77.8	82.6
	(0.1)	(1.1)	(0.1)	(0.4)	(0.5)	(0.4)
FedAvg	95.9	62.9	95.6	82.3	76.9	82.7
	(0.2)	(1.5)	(0.3)	(0.4)	(0.5)	(0.6)
FedProx	95.8	63.1	95.6	82.3	76.6	82.7
	(0.2)	(1.6)	(0.3)	(0.4)	(0.6)	(0.6)
FedBN	96.6	71.0	97.0	83.2	78.3	85.2
	(0.1)	(0.3)	(0.3)	(0.4)	(0.7)	(0.4)
pFedSD	95.7	64.8	95.4	81.6	80.7	83.6
	(0.1)	(0.4)	(0.2)	(0.5)	(0.3)	(0.3)
FedSSD	**96.9**	61.9	96.7	82.2	79.1	83.4
	(0.2)	(1.3)	(0.2)	(0.7)	(0.7)	(0.6)
FedWon	96.5	**71.3**	96.3	85.4	77.7	85.4
	(0.1)	**(0.3)**	(0.2)	(0.3)	(0.7)	(0.3)
pFedDB	96.7	68.6	**97.0**	**86.4**	**85.2**	**86.8**
	(0.1)	(0.1)	**(0.1)**	**(0.8)**	**(0.1)**	**(0.3)**

Table 4: Office-Caltech-10 multi-domain classification accuracy (%). Mean (top) and standard deviation (bottom).

Method	A	C	D	W	Avg
AlexNet	54.9	40.2	78.7	86.4	65.1
	(1.5)	(1.6)	(1.3)	(2.4)	(1.7)
FedAvg	54.1	44.8	66.9	85.1	62.7
	(1.1)	(1.0)	(1.5)	(2.9)	(1.6)
FedProx	54.2	44.5	65.0	84.4	62.0
	(2.5)	(0.5)	(3.6)	(1.7)	(2.1)
FedBN	63.0	45.3	83.1	90.5	70.5
	(1.6)	(1.5)	(2.5)	(2.8)	(2.0)
pFedSD	64.1	45.7	90.6	90.4	72.7
	(1.0)	(0.5)	(2.3)	(1.3)	(0.6)
FedSSD	59.8	46.0	87.5	87.1	70.1
	(1.1)	(1.6)	(2.0)	(2.5)	(0.4)
FedWon	**65.1**	49.2	90.6	83.7	72.2
	(2.2)	(2.1)	(2.8)	(4.5)	(2.3)
pFedDB	62.5	**50.1**	**90.6**	**90.5**	**73.4**
	(1.1)	**(0.6)**	**(1.5)**	**(2.3)**	**(1.1)**

Table 5: DomainNet multi-domain classification accuracy (%). Mean (top) and standard deviation (bottom).

Method	C	I	P	Q	R	S	Avg
AlexNet	41.0	23.8	36.2	73.1	48.5	34.0	42.8
	(0.9)	(1.2)	(2.7)	(0.9)	(1.9)	(1.1)	(1.5)
FedAvg	48.8	24.9	36.0	56.1	46.3	36.6	41.5
	(1.9)	(0.7)	(1.1)	(1.6)	(1.4)	(2.5)	(1.5)
FedProx	48.9	24.9	36.6	54.4	47.8	36.9	41.6
	(0.8)	(1.0)	(1.8)	(3.1)	(0.8)	(2.1)	(1.6)
FedBN	51.2	26.8	41.5	71.3	54.8	42.1	48.0
	(1.4)	(0.5)	(1.4)	(0.7)	(0.8)	(1.3)	(1.0)
pFedSD	49.4	26.9	39.9	70.5	53.3	37.4	46.2
	(2.0)	(0.6)	(1.3)	(1.1)	(1.1)	(2.3)	(0.9)
FedSSD	51.0	26.6	37.2	69.5	50.8	36.8	45.3
	(1.9)	(0.9)	(1.3)	(2.2)	(0.5)	(1.6)	(0.6)
FedWon	**54.7**	**26.9**	40.0	68.3	54.0	48.9	48.8
	(1.3)	**(0.7)**	(1.0)	(1.3)	(0.7)	(1.6)	(0.6)
pFedDB	48.5	25.0	**44.1**	**74.0**	**60.2**	**49.9**	**50.3**
	(0.7)	(0.4)	**(0.9)**	**(0.4)**	**(0.8)**	**(0.4)**	**(0.3)**

References

Demner-Fushman, D.; Kohli, M. D.; Rosenman, M. B.; Shooshan, S. E.; Rodriguez, L.; Antani, S.; Thoma, G. R.; and McDonald, C. J. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2): 304–310.

Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilcus, S.; Chute, C.; Marklund, H.; Haghgoo, B.; Ball, R.; Shpanskaya, K.; et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 590–597.

Nguyen, H. Q.; Lam, K.; Le, L. T.; Pham, H. H.; Tran, D. Q.; Nguyen, D. B.; Le, D. D.; Pham, C. M.; Tong, H. T.; Dinh, D. H.; et al. 2022. VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations. *Scientific Data*, 9(1): 429.

Shih, G.; Wu, C. C.; Halabi, S. S.; Kohli, M. D.; Prevedello, L. M.; Cook, T. S.; Sharma, A.; Amorosa, J. K.; Arteaga, V.; Galperin-Aizenberg, M.; et al. 2019. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*, 1(1): e180041.