FedSwitch: From Standard to Balanced Aggregation via Private Histogram Estimation in Federated Medical Image Classification

Paolo Cacace[1,2✉], Riccardo Taiello[1], Andrea Protani[1,3], Marc Molina Van den Bosch[1,3], Diogo Reis Santos[1], Pierpaolo Brutti[2], and Luigi Serio[1]

[1] CERN, Geneva, Switzerland
paolo.cacace@cern.ch
[2] DIAG Department, Sapienza University, Rome, Italy
[3] BCN MedTech, Department of Engineering, Universitat Pompeu Fabra, Barcelona, Spain

Abstract. Federated Learning (FL) enables collaborative model training across institutions without sharing raw data. Performance degrades, however, when label distributions are imbalanced across sites, a common scenario in medical settings where disease prevalence could vary drastically across institutions. Existing distribution-aware methods rely on label distribution statistics that clients cannot share in real-world federated settings due to privacy and regulatory constraints. We propose FedSwitch, a model-agnostic FL framework that estimates the global label histogram, under client-level ε-differential privacy, and uses the estimate to switch from FedAvg to distribution-aware aggregation. Each client adds partial discrete Laplace noise to its bounded label counts and secret-shares the result via Shamir's scheme; designated decryptors reconstruct the noisy aggregate without ever observing individual contributions. The optimal switching round can be estimated via a closed-form variance decomposition that accounts for bounded contributions, client sampling, and differential-privacy noise components. We evaluate FedSwitch on medical imaging benchmarks under realistic heterogeneous settings, demonstrating consistent improvements over standard FL aggregation with negligible communication overhead.

Keywords: Federated Learning · Privacy Enhancing Technologies · Non-IID Data Distribution

1 Introduction

Federated Learning (FL) [14] enables multi-institutional model training while keeping patient data on-site, a critical requirement under regulatory frameworks such as GDPR and HIPAA. In multi-class classification tasks, statistical heterogeneity from non-IID label distributions remains a major challenge [10]. FedAvg [14] weights client updates according to local dataset size, implicitly assuming comparable label distributions. In medical contexts, disease prevalence could

vary substantially across institutions, biasing the model toward dominant conditions, reducing performance on rare yet clinically relevant pathologies. Several methods address this by exploiting label-distribution information.

Client-side approaches [22,20] address label distribution locally while retaining FedAvg style aggregation. Optimization-based methods, such as [12], reduce client drift, but don't target label skew directly.

Server-side reweighting approaches [11] tackle label-skew by accessing global label statistics, but require clients to disclose sensitive information. Extensions adding cryptographic protection [2], protect clients' individual contributions, yet provide no formal differential privacy (DP) [5] guarantees and are likewise evaluated under full participation. A naive approach could be estimating the histogram with off-the-shelf federated analytics [3], combining individual contribution protection with differential privacy. This becomes problematic with partial client participation: SecAgg [4] exposes only a per-round histogram aggregate over the sampled clients, so achieving broad coverage requires releasing such a DP aggregate at every round, and the associated privacy cost composes across rounds. The injected noise then grows with the number of releases and quickly makes the estimated histogram unusable.

We propose FedSwitch, a model-agnostic FL framework that retains the effectiveness of *server-side* distribution-aware aggregation while guaranteeing client-level DP for the disclosed label statistics. Leveraging the label-aware, inverse-frequency weights of [11], our contribution resides in a private estimation protocol that makes server-side reweighting deployable under partial participation and without a trusted aggregator. Each client bounds its contribution and injects partial discrete Laplace noise [3], then secret-shares its noisy counts via Shamir's scheme [16]. This pipeline allows shares to accumulate in the server during a warm-up phase that runs in parallel with standard FedAvg aggregation; the histogram is recovered through a single one-shot reconstruction at round $R^\star$ rather than refreshed per round, incurring no privacy composition, in contrast to the naive scheme above. Leveraging a closed-form variance decomposition formulation (Eq. (3)), the optimal round $R^\star$ at which to trigger the distribution-aware aggregation can be estimated, replacing FedAvg's dataset-size weights with inverse-frequency weights derived from the reconstructed histogram. The proposed pipeline allows estimating the private label distribution *once*, reducing the total privacy expenditure to a single ε-DP query, avoiding composition over rounds.

Our main contributions are: *(i)* A privacy-preserving histogram-estimation protocol that combines distributed DP noise generation with Shamir secret sharing, achieving client-level ε-DP without a trusted aggregator. *(ii)* A closed-form uncertainty estimation criterion that identifies the estimated histogram drift, accounting for contribution bounding, client sampling, and DP noise. *(iii)* Evaluation on three medical-imaging benchmarks (HAM10000, BloodMNIST, OrganMNIST3D) showing that, under severe heterogeneity, existing methods fail to capture rare pathologies, while FedSwitch achieves > 0.45 F1 on the rarest class of HAM10000 ($< 2\%$ prevalence).

Code is available at GitHub Repo.

2 Problem statement

We consider a k-class classification problem in a federated setting with N clients $\mathcal{C} = \{P_1, \ldots, P_N\}$ and a server $\mathcal{S}$. Each client P_i holds a private dataset $\mathcal{D}_i$ of size n_i with label-count vector $\mathbf{y}_i = (c_{i,1}, \ldots, c_{i,k}) \in \mathbb{Z}^k_{\geq 0}$. At each round $r \in [1, \ldots, T]$ the server samples a subset $\mathcal{P}^r \subseteq \mathcal{C}$. Every selected client trains the current global model θ^r to produce an update $\Delta^r_i = \theta^r_i - \theta^r$, and the server aggregates:

$$\theta^{r+1} = \theta^r + \sum_{P_i \in \mathcal{P}^r} w^r_i \, \Delta^r_i, \qquad w^r_i \geq 0, \;\; \textstyle\sum_{P_i \in \mathcal{P}^r} w^r_i = 1. \tag{1}$$

Standard FedAvg [14] sets $w^r_i = n_i / \sum_{P_j \in \mathcal{P}^r} n_j$, weighting by dataset size. Under class imbalance, this biases the model toward prevalent classes. Distribution-aware weights can mitigate this effect but require knowledge of the client's label histogram $\mathbf{y} = \sum_i \mathbf{y}_i$, which clients are unwilling to disclose. FedSwitch addresses this problem by estimating $\mathbf{y}$ under ε-*DP* at client level, then using the estimate to set the weights w^r_i, without revealing any individual $\mathbf{y}_i$.

Threat model. We adopt the standard secure-aggregation threat model [4], where clients and a designated subset of M honest-but-curious *decryptors* $\mathcal{D} = \{D_1, \ldots, D_M\} \subseteq \mathcal{C}$ participate [13], and the server may collude with up to $M - t - 1$ decryptors, with $t = \lceil 2M/3 \rceil$. Decryptors act as neutral intermediaries: each aggregates encrypted shares, while remaining unable to decrypt individual contributions independently. Only when at least t decryptors combine their accumulated shares, the aggregated histogram can be recovered, ensuring that neither the server nor any single decryptor observes individual label counts.

Privacy goal. We provide client-level ε-DP for the histogram: each client's entire label-count vector is protected regardless of the number of samples it holds. The guarantee covers the histogram estimation. Model update privacy is orthogonal and can be addressed by local DP [17], distributed DP [9], DP-SGD [1] and SecAgg [4].

3 Method

FedSwitch operates in three stages (Fig. 1). During the *warm-up* (rounds 1 to $R^\star - 1$), clients generate and distribute secret shares of their noisy label counts, while standard FedAvg training runs in parallel. At round $R^\star$, the server triggers *histogram reconstruction*: decryptors receive the accumulated shares from the server and combine them to enable the revealing of the aggregated label histogram under ε-DP. From round $R^\star + 1$ onward, the server performs *distribution-aware aggregation*, the core optimization of FedSwitch, using inverse-frequency weights derived from the reconstructed histogram.

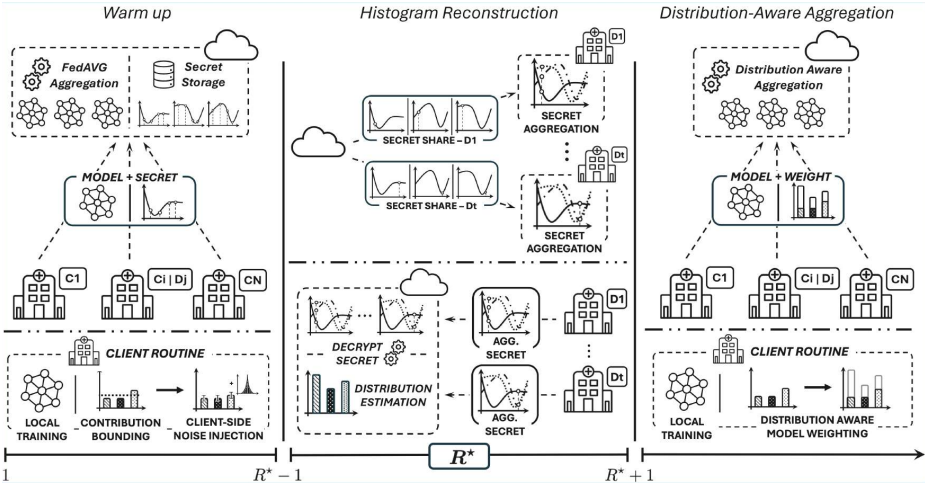

Fig. 1. Method overview. *Left:* during warm-up, each client trains locally and secret-shares its noisy label counts to the server. *Center:* at round $R^\star$, decryptors combine shares and return them to the server, allowing reconstruction of the DP histogram. *Right:* from round $R^\star+1$ onward, clients send model updates with their distribution-aware weight, enabling the server to perform distribution-aware aggregation.

Warm-up: training and share accumulation (rounds 1 ***to*** $R^\star-1$***).*** The warm-up serves primarily for clients' *secret share accumulation*: with each round increasing the observed client fraction ρ at no extra privacy cost. The server selects m clients per round. Each selected client P_i performs two tasks in parallel: (i) one-time generation of secret shares for the histogram protocol below, and (ii) local training and upload of Δ_i^r. In this phase, the server aggregates the clients' updates under standard FedAvg weights ($w_i^r = n_i / \sum_j n_j$) and buffers incoming secret shares across rounds.

Contribution bounding. Each client P_i draws C samples uniformly with replacement from $\mathcal{D}_i$ and computes a contribution-bounded label-count vector $\tilde{\mathbf{y}}_i \in \mathbb{Z}^k_{\geq 0}$ with $\|\tilde{\mathbf{y}}_i\|_1 = C$, bounding the ℓ_1-sensitivity to exactly C regardless of n_i. Since a client must hold at least C samples to participate, C implicitly defines a minimum-participation threshold.

Client-side noise. To achieve ε-DP, discrete Laplace $\mathrm{DLap}(C/\varepsilon)$ noise is added to the aggregated histogram. Rather than relying on a trusted party, we distribute the noise generation across the seen clients at round $R^\star$ ($|\mathcal{A}| = \lceil \rho N \rceil$) using the Pólya decomposition of [3]: each client P_i independently samples, for every label ℓ:

$$\eta_{i,\ell} = X^+_{i,\ell} - X^-_{i,\ell}, \qquad X^\pm_{i,\ell} \sim \text{Pólya}\big(\tfrac{1}{|\mathcal{A}|},\, e^{-\varepsilon/C}\big). \tag{2}$$

When all $|\mathcal{A}|$ clients contribute, the partial terms sum to a full discrete Laplace: $\sum_{i=1}^{|\mathcal{A}|} \eta_{i,\ell} \sim \mathrm{DLap}(C/\varepsilon)$, guaranteeing ε-DP for the aggregated histogram.

Secret sharing [16]. Each client secret-shares its *noisy* bounded count $\tilde{c}_{i,\ell} + \eta_{i,\ell}$ via Shamir's scheme: P_i constructs a degree-$(t-1)$ polynomial with free term $\tilde{c}_{i,\ell} + \eta_{i,\ell}$ and sends one share per decryptor to the server. The secret shares are crafted in such a way that fewer than t shares reveal nothing about the secret. Client-share secret generation is performed *once*. Per-client cost is $O(kM)$ communication and $O(kM^2)$ computation, reducible to $O(k)$ via packed secret sharing [6].

Histogram reconstruction (round $R^\star$). At round $R^\star$, the observed set $\mathcal{A}$ has reached size $\lceil \rho N \rceil$. Each decryptor D_j sums the secret shares received from the server: $\bar{s}_\ell(j) = \sum_{P_i \in \mathcal{A}} s_{i,\ell}(j) \bmod p$ and sends it back to the server. With at least t responses, the server can reconstruct $\hat{y}_\ell = \sum_j \lambda_j \bar{s}_\ell(j) \bmod p = \sum_{P_i \in \mathcal{A}} (\tilde{c}_{i,\ell} + \eta_{i,\ell})$ via Lagrange interpolation, i.e. the true aggregate plus distributed noise. If fewer than t decryptors respond, reconstruction is deferred and FedSwitch continues with FedAvg weights. Each decryptor stores $O(k)$ shares per client and performs $O(k|\mathcal{A}|)$ additions at round $R^\star$. This cost is incurred once and is negligible compared to model updates. The server then broadcasts $\hat{\mathbf{y}}$ to all participants.

Distribution-aware aggregation (rounds $R^\star{+}1$ to T). Each selected client locally computes an aggregation score comparing the aggregated histogram and its own:

$$\phi_i = \min\left(1,\ \sum_{\ell=1}^{k} \frac{c_{i,\ell}}{\max(\delta,\ \hat{y}_\ell)}\right),$$

where $\delta = 1$ prevents division by zero. The score ϕ_i upweights clients holding globally rare classes. The aggregation weights are $w_i^r = \phi_i / \sum_{P_j \in \mathcal{P}^r} \phi_j$, extending the label-aware concept of [11] to a privacy-preserving setting. Since ϕ_i is a single scalar (analogous to n_i in FedAvg), standard secure aggregation [4] protects it: the server only observes the aggregate $\sum_j \phi_j$ and each client computes w_i^r locally, eliminating per-client leakage at the cost of one additional scalar per round.

Histogram uncertainty estimation. The estimate $\hat{\mathbf{y}}$ combines three error sources: partial client coverage, contribution bounding, and DP noise. For a fixed label ℓ, let $q_{i,\ell}$ be the bounded local proportion at client P_i (approximating the client distribution using C samples), $\bar{q}_\ell = \frac{1}{N}\sum_i q_{i,\ell}$ the mean across clients, and $S_\ell^2 = \mathrm{Var}(q_{i,\ell})$ the between-client variance. The ratio $\kappa_\ell = S_\ell^2 / [\bar{q}_\ell(1-\bar{q}_\ell)] \in [0,1]$ measures how much of the total variability is due to cross-site heterogeneity rather than within-client sampling. With $|\mathcal{A}| = \rho N$ observed clients the estimation variance decomposes as:

$$\mathrm{Var}(\hat{p}_\ell - \bar{q}_\ell) = \frac{\bar{q}_\ell(1-\bar{q}_\ell)}{|\mathcal{A}|} \left[\underbrace{\frac{1-\kappa_\ell}{C}}_{\substack{\text{contribution}\\ \text{bounding}}} + \underbrace{\kappa_\ell(1-\rho)}_{\text{client sampling}} \right] + \underbrace{\sigma^2_{\mathrm{DP}}}_{\text{DP noise}}, \qquad (3)$$

where $\sigma^2_{\mathrm{DP}} = \rho \cdot 2\beta/(1-\beta)^2$, $\beta = e^{-\varepsilon/C}$, is the variance from the distributed Pólya noise (Eq. (2)). All three terms shrink as coverage ρ grows. In our setting, the server uniformly samples m clients per round with replacement across rounds. The reconstruction is triggered at:

$$R^\star = \left\lfloor \frac{\log(1-\rho)}{\log(1-m/N)} \right\rfloor,$$

which corresponds to the earliest round at which the coverage ρ is achieved.

Practical implications. While $\bar{q}_\ell$ is a fixed parameter that can be estimated, particularly in medical settings, both C and ρ remain tunable. Setting the C value involves a trade-off: a larger C reduces the contribution-bounding variance (first term of Eq. (3)) but amplifies DP noise ($\beta \to 1$) and excludes more sites, while a small C increases client participation, estimation noise and exposes the aggregation model to numerical instability from extremely small sites [8]. Higher coverage target ρ shrinks both sampling and DP terms at the cost of a later switch round $R^\star$, extending the FedAvg warm-up whose aggregation is biased under label skew. In our experiments, we set $C = \min n_i$ so that no client is excluded. In a realistic deployment setting, where individual dataset sizes are not disclosed, $\min n_i$ can be estimated via a one-shot distributed DP query over dataset sizes [15], and C can then be set.

4 Experiments

We evaluate FedSwitch on three medical imaging classification benchmarks that span 2D and 3D modalities: HAM10000 [18], BloodMNIST [21], and OrganMNIST3D [21]. All experiments compare FedSwitch against FedAvg [14], FedProx [12], FedIIC [20], and FedLC [22] using the same uniform client sampling strategy.

Implementation details. We use Adam with learning rate 10^{-4} and weight decay 10^{-4}. Each client i trains for $E = \lfloor C/B \rfloor$ local optimization steps per round, where B is the batch size, and $C = \min n_i$ for each client. We use a fixed number of steps rather than epochs: as shown by Wang et al. [19], preventing objective inconsistencies. For FedProx we set the proximal coefficient $\mu = 0.01$. The histogram estimation uses $\varepsilon = 1$ and target coverage $\rho = 0.9$ for all datasets. Hyperparameters are kept constant across methods to ensure a fair comparison. All experiments are repeated over 5 random seeds, and results are reported as macro-F1 (mean ± std).

Datasets. **HAM10000 [18]** comprises 10,015 dermoscopic RGB images (600× 450 px, resized to 224×224) across 7 diagnostic categories. The dataset is heavily imbalanced: the majority class (nv) accounts for 66.9% of images, the rarest (df) only 1.2%. We use $N = 100$ clients, $m = 10$ per round, and a ResNet-50 backbone (the larger input resolution requires a deeper network than MedMNIST). We set $C = 50$, $B = 16$, $E = 3$, $R^\star = 21$.

BloodMNIST [21] contains 17,092 microscopic blood-cell images (28×28 px, 8 classes). The dataset is mildly imbalanced: the majority class (neutrophil) accounts for 19.5% of samples, while the rarest class (lymphocyte) represents 7.1%. We merge the official training and validation splits (13,671 images) and distribute them among $N=100$ clients with $m=10$ per round, using a ResNet-18 backbone as in [21]. We set $C=70$, $B=32$, $E=2$, $R^\star=21$.

OrganMNIST3D [21] contains 1,742 single-channel CT volumes of size 28^3, spanning 11 anatomical organ classes. To reduce label redundancy, left-right paired organs are merged, resulting in 8 classes. In this setup, the dataset is mildly imbalanced: the majority class (femur) accounts for 22.8% of volumes, whereas the rarest class (liver) accounts for 4.7%. The official training and validation splits are partitioned across $N = 50$ clients, with $m = 10$ clients sampled per communication round. Model training employs a 3D ResNet-18 following [21]. We set the hyperparameters as $C = 12$, $B = 8$, $E = 1$, and $R^\star = 10$.

Federated partitioning. To simulate realistic heterogeneity, we combine two complementary sources of non-IID-ness in all synthetic partitions.

Quantity skew (fixed). Client dataset sizes are sampled using a Min-Dirichlet [7] draw with a concentration $\alpha_s=0.5$ and a minimum allocation of at least half the uniform dataset samples share, inducing moderate inter-client size variability.

Label skew (swept). Following the distribution-based partitioning of [22], each client i draws class proportions $\pi_i \sim \mathrm{Dir}(\alpha_\ell \mathbf{1}_k)$ and samples its labels accordingly. We denote this scheme $\mathrm{D}(\alpha_\ell)$, where D represents a Dirichlet-based partitioning: smaller α_ℓ yields highly concentrated class supports (with missing class coverage at some clients), while larger values approach uniform mixtures. We sweep $\alpha_\ell \in \{0.1, 0.5, 1.0\}$ to cover severe-to-mild label heterogeneity.

Quantitative results. Table 1 reports macro-F1 across all dataset-skew combinations; centralized upper bounds are shown alongside each dataset name.

The largest gains appear on **HAM10000**, whose heavy intrinsic imbalance (66.9% majority class) amplifies the federated label skew and the effect of the distribution-aware weighting scheme. At D(0.1), every baseline collapses toward majority-class prediction (macro-F1 ≤ 0.15), whereas FedSwitch ($\varepsilon = 2$) reaches 0.61 (+0.46). This advantage is consistent even at D(1.0), where FedSwitch achieves 0.73, within 0.09 of the centralized bound, while the best baseline (FedLC, 0.37) remains 0.36 lower, confirming that client-side corrections alone cannot compensate for a skewed global distribution.

On **BloodMNIST** and **OrganMNIST3D**, which both present more balanced overall label class distribution, the performance improvements are reduced: FedSwitch improves by +0.06 and +0.02 at D(0.1), and at D(1.0) all methods converge to within ≤ 0.04 of each other, with the residual gap to the centralized bound attributable to the federated setting itself.

Across settings, $\varepsilon=2$ matches or slightly exceeds $\varepsilon=1$, consistent with lower DP noise yielding a more accurate histogram estimation.

Table 1. Macro-F1 across datasets and label heterogeneity levels. Best federated results in **bold**. Standard deviations over 5 seeds are shown as subscripts. The centralized upper bound is reported next to each dataset name. **HAM** refers to **HAM10000**, **Blood** refers to **BloodMNIST** and **Organ** refers to **OrganMNIST3D**. For FedSwitch, experiments are performed with a privacy budget $\varepsilon \in \{1, 2\}$.

Dataset *Centralized F1*	α_ℓ	**Baselines**				**FedSwitch**	
		FedAvg	FedProx	FedIIC	FedLC	$\varepsilon=1$	$\varepsilon=2$
HAM (2D) *0.82*$_{\pm 0.02}$	0.1	$0.13_{\pm.01}$	$0.15_{\pm.02}$	$0.14_{\pm.01}$	$0.15_{\pm.02}$	$0.56_{\pm.04}$	$\mathbf{0.61}_{\pm.03}$
	0.5	$0.29_{\pm.02}$	$0.28_{\pm.01}$	$0.28_{\pm.02}$	$0.29_{\pm.01}$	$\mathbf{0.73}_{\pm.02}$	$0.72_{\pm.01}$
	1.0	$0.35_{\pm.04}$	$0.18_{\pm.10}$	$0.36_{\pm.01}$	$0.37_{\pm.02}$	$0.72_{\pm.02}$	$\mathbf{0.73}_{\pm.01}$
Blood (2D) *0.96*$_{\pm 0.01}$	0.1	$0.63_{\pm.01}$	$0.63_{\pm.01}$	$0.52_{\pm.03}$	$0.63_{\pm.02}$	$0.64_{\pm.01}$	$\mathbf{0.69}_{\pm.03}$
	0.5	$0.81_{\pm.01}$	$0.81_{\pm.01}$	$0.78_{\pm.02}$	$0.80_{\pm.03}$	$0.83_{\pm.01}$	$\mathbf{0.83}_{\pm.01}$
	1.0	$0.87_{\pm.01}$	$0.87_{\pm.01}$	$0.86_{\pm.02}$	$0.85_{\pm.02}$	$0.87_{\pm.01}$	$\mathbf{0.87}_{\pm.02}$
Organ (3D) *0.95*$_{\pm 0.01}$	0.1	$0.73_{\pm.01}$	$0.73_{\pm.01}$	$0.77_{\pm.02}$	$0.84_{\pm.02}$	$0.85_{\pm.01}$	$\mathbf{0.86}_{\pm.01}$
	0.5	$0.87_{\pm.01}$	$0.87_{\pm.01}$	$0.86_{\pm.01}$	$0.87_{\pm.02}$	$0.88_{\pm.01}$	$\mathbf{0.88}_{\pm.01}$
	1.0	$0.86_{\pm.02}$	$0.85_{\pm.02}$	$0.88_{\pm.02}$	$\mathbf{0.89}_{\pm.01}$	$0.88_{\pm.02}$	$0.88_{\pm.02}$

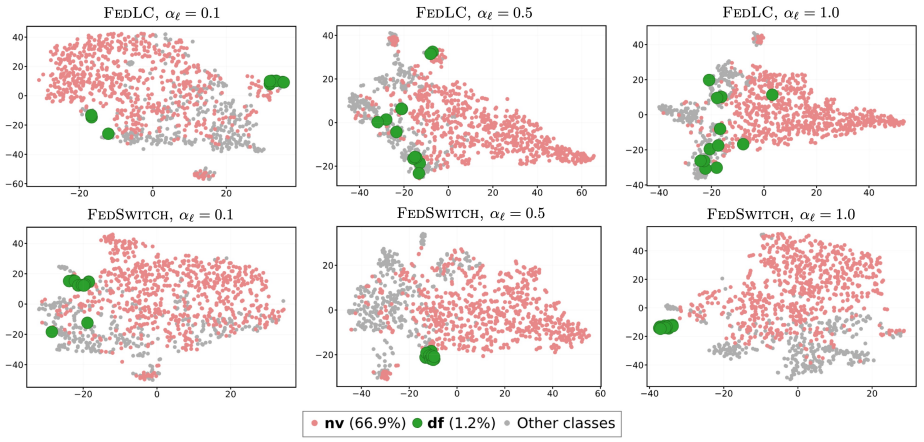

Fig. 2. t-SNE of feature embeddings from the last hidden layer on HAM10000 for FedLC (top) and FedSwitch with $\varepsilon=1$ (bottom) under increasing label heterogeneity. The majority class **nv** (66.9%, red) and rarest class **df** (1.2%, green) are highlighted.

Qualitative analysis. Figure 2 shows t-SNE embeddings of penultimate-layer features on HAM10000 for FedLC (top performance) and FedSwitch at $\varepsilon=1$ (bottom). Under severe skew ($\alpha_\ell = 0.1$), FedLC's **df** samples (1.2%) scatter across the embedding space, whereas FedSwitch produces a visibly tighter grouping that consolidates into a well-separated cluster as α_ℓ increases, suggesting that inverse-frequency weighting allocates representational capacity to rare classes even when few clients contribute to them.

Per-class scores confirm this: FedLC achieves < 0.1 F1 on **df** at every heterogeneity level, indicating complete failure on the rarest class. FedSwitch averages 0.46, 0.58, and 0.68 at D(0.1), D(0.5), and D(1.0) respectively.

5 Conclusion and Future Work

This work introduced FedSwitch, a privacy-preserving framework for federated classification on class-imbalanced medical data. The method works by estimating the global label histogram once under client-level ε-DP to enable distribution-aware aggregation without revealing client datasets' distributions. A variance-based trigger allows for determining the optimal round to switch from FedAvg to inverse-frequency weighting. Experiments across 2D and 3D medical imaging benchmarks demonstrate consistent improvements over strong baselines, particularly under severe label heterogeneity, with negligible communication overhead. Future work will focus on integrating model-update privacy mechanisms, enabling drift-aware histogram re-estimation, automating contribution-bound selection, and extending the framework to broader medical tasks.

Acknowledgments. This project is supported by the CERN Knowledge Transfer Fund for Medical Applications. The authors also thank Lisa Guzzi for the insightful discussions during the writing of this paper.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L.: Deep learning with differential privacy. In: CCS. pp. 308–318. ACM (2016)
2. Akram, A., Pham, H.N.H., Önen, M., Gritti, C.: SAAFL: Secure aggregation for label-aware federated learning. In: IFIP SEC. IFIP AICT, Springer (2025)
3. Bagdasaryan, E., Kairouz, P., Mellem, S., Gascón, A., Bonawitz, K., Estrin, D., Gruteser, M.: Towards sparse federated analytics: Location heatmaps under distributed differential privacy with secure aggregation. PoPETs **2022**(4), 162–182 (2022)
4. Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H.B., Patel, S., Ramage, D., Segal, A., Seth, K.: Practical secure aggregation for privacy-preserving machine learning. In: CCS. pp. 1175–1191 (2017)
5. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Halevi, S., Rabin, T. (eds.) TCC 2006. LNCS, vol. 3876, pp. 265–284. Springer, Heidelberg (2006)
6. Franklin, M., Yung, M.: Communication complexity of secure computation. In: STOC. pp. 699–710. ACM (1992)
7. Jimenez G., D.M., Anagnostopoulos, A., Chatzigiannakis, I., Vitaletti, A.: FedArtML: A tool to facilitate the generation of non-IID datasets in a controlled way to support federated learning research. IEEE Access **12**, 81004–81016 (2024)

8. Jimenez G., D.M., Solans, D., Heikkilä, M., Vitaletti, A., Kourtellis, N., Anagnostopoulos, A., Chatzigiannakis, I.: Non-IID data in federated learning: A survey with taxonomy, metrics, methods, frameworks and future directions. arXiv preprint arXiv:2411.12377 (2024)
9. Kairouz, P., Liu, Z., Steinke, T.: The distributed discrete gaussian mechanism for federated learning with secure aggregation. In: ICML. PMLR, vol. 139, pp. 5201–5212 (2021)
10. Kairouz, P., McMahan, H.B., Avent, B., et al.: Advances and open problems in federated learning. Found. Trends Mach. Learn. **14**(1–2), 1–210 (2021)
11. Khalil, A., Wainakh, A., Zimmer, E., Parra-Arnau, J., Fernandez Anta, A., Meuser, T., Steinmetz, R.: Label-aware aggregation for improved federated learning. In: IEEE FMEC. pp. 216–223 (2023)
12. Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V.: Federated optimization in heterogeneous networks. In: MLSys (2020)
13. Ma, Y., Woods, J., Angel, S., Polychroniadou, A., Rabin, T.: Flamingo: Multi-round single-server secure aggregation with applications to private federated learning. In: IEEE S&P. pp. 477–496 (2023)
14. McMahan, B., Moore, E., Ramage, D., Hampson, S., Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: AISTATS. pp. 1273–1282. PMLR (2017)
15. Pillutla, K., Laguel, Y., Malick, J., Harchaoui, Z.: Differentially private federated quantiles with the distributed discrete gaussian mechanism. In: Workshop on Federated Learning, NeurIPS (2022)
16. Shamir, A.: How to share a secret. Commun. ACM **22**(11), 612–613 (1979)
17. Truex, S., Liu, L., Chow, K.H., Gursoy, M.E., Wei, W.: LDP-Fed: Federated learning with local differential privacy. In: EdgeSys. pp. 61–66. ACM (2020)
18. Tschandl, P., Rosendahl, C., Kittler, H.: The HAM10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data **5**, 180161 (2018)
19. Wang, J., Liu, Q., Liang, H., Joshi, G., Poor, H.V.: Tackling the objective inconsistency problem in heterogeneous federated optimization. In: NeurIPS. vol. 33, pp. 7611–7623 (2020)
20. Wu, N., Yu, L., Yang, X., Cheng, K.T., Yan, Z.: FedIIC: Towards robust federated learning for class-imbalanced medical image classification. In: MICCAI. LNCS, vol. 14221, pp. 692–702. Springer (2023)
21. Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B.: MedMNIST v2: A large-scale lightweight benchmark for 2D and 3D biomedical image classification. Scientific Data **10**, 41 (2023)
22. Zhang, J., Li, Z., Li, B., Xu, J., Wu, S., Ding, S., Wu, C.: Federated learning with label distribution skew via logits calibration. In: ICML. pp. 26311–26329. PMLR (2022)