Label-Aware Federated Learning Improves ECG Classification in Heterogeneous Hospital Settings

Tizian C. Dege[1], Christoph Hoog Antink[1]
[1]KIS*MED - AI Systems in Medicine,
Technical University of Darmstadt, Darmstadt, Germany

Abstract

Federated Learning (FL) enables multiple clients to jointly train a model without sharing raw data. However, traditional FL algorithms such as FedAvg struggle with heterogeneous (non-IID) data, letting the largest client dominate the shared model. We analyze three clients from the PhysioNet/CinC 2020 Challenge dataset – CPSC (n=4108), Georgia (n=5489), and PTB-XL (n=20517) – under three ECG classification settings: (i) decentralized, client-only training; (ii) a central baseline with all data pooled; and (iii) a federated approach sharing only model updates. For (iii), we compare FedAvg and FedProx, which weight clients by sample size, against their label-aware counterparts FedLA and FedProx+LA, which weight by label share instead. All settings use 5-fold cross-validation across three padding strategies. Our results show that the best aggregation method depends on client size and label skew: FedAvg is the strongest choice for the majority client (PTB-XL), while the label-aware methods are the strongest choice for the minority, most label-skewed client (CPSC). The Georgia client gains nothing from federation under any aggregation method, so the benefit of federation is distributed unevenly across clients. We find this benefit to be largest where local data is scarcest, with CPSC gaining 0.079 AUROC over local-only training under FedProx+LA, compared to 0.006 for PTB-XL under FedAvg.

1. Introduction

Cardiovascular Diseases (CVDs) remain the leading cause of death worldwide, responsible for an estimated 20.5 million deaths in 2021, roughly 80% of which occurred in low- and middle-income countries [1]. The 12-lead ECG remains the most widely used tool for detecting and monitoring CVDs, and deep neural networks now classify multiple concurrent abnormalities directly from raw ECG waveforms. ECG data collected across different hospitals is nevertheless inherently heterogeneous, for two distinct reasons. First, acquisition hardware, sampling protocols, and recording duration differ by manufacturer and site. Second, independently of any single institution's practice, patient populations differ in age, sex, and region-specific environmental exposure. Together, these data- and population-level differences mean that the diagnoses observed at one site are rarely representative of those observed elsewhere, so training assumptions that treat hospital data as identically distributed are unlikely to hold in practice.

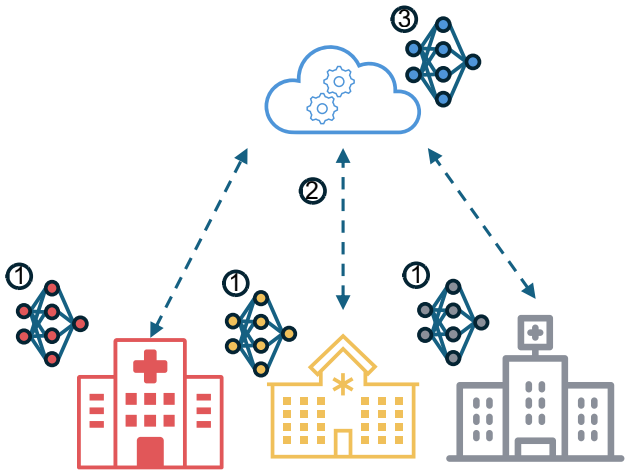

Figure 1. In a FL scenario, each client first trains a local model (1). Model updates are exchanged with a central server (2), which aggregates them into a global model (3); the updated global model is then returned to each client for another round of local training, and the process repeats over multiple communication rounds.

Moreover, ECG recordings are protected health information, and institutional and regulatory constraints frequently prevent hospitals from sharing raw signals across sites. Federated Learning (FL) offers a way to train a shared model under these constraints: each site trains locally and transmits only model updates, which a server aggregates, so that raw waveforms never leave the institution [2]. This makes FL an attractive tool for multi-institution ECG classification, but the heterogeneity described above turns training itself into a statistical chal-

lenge: an aggregation rule that is naive to differences in each site's device characteristics and diagnosis distribution risks a shared model that reflects the largest or most homogeneous site rather than the federation as a whole. This paper studies how the choice of federated aggregation algorithm affects multi-label 12-lead ECG classification across three heterogeneous sources, with particular attention to per-client performance under label imbalance.

Related Work: Multiple frameworks exist for decentralized training. Federated Averaging (FedAvg) remains the original and most widely used baseline, in which clients run local stochastic gradient descent and the server aggregates their updates using weights proportional to each client's number of training samples [2]. When client data are heterogeneous, this sample-size weighting lets the largest client dominate the aggregate. FedProx addresses client drift under heterogeneity by adding an ℓ_2 proximal term to each client's local objective, penalizing divergence from the current global model while retaining FedAvg's sample-size aggregation weights [3]. FedLA instead modifies the aggregation step itself, replacing sample-size weights with weights derived from each client's share of the global label distribution, so that clients holding a disproportionate share of a given label's examples are upweighted independently of their raw dataset size [4]. FedProx+LA combines both ideas, applying FedLA's label-aware weights on top of FedProx's proximal term. We benchmark these four aggregation strategies on the PhysioNet/CinC 2020 Challenge dataset, whose multi-source design makes this kind of heterogeneity inevitable [5].

2. Methods

Dataset and clients We use the PhysioNet 2020 Challenge training data [5] and treat three of its sources as federated clients: CPSC, Georgia, and PTB-XL. Each recording is a 12-lead waveform annotated with one or more SNOMED CT diagnosis codes. Rather than the full officially-scored label set, we target the five most frequent officially-scored diagnoses computed over the pooled data and applied identically to all clients: sinus rhythm (SR), left axis deviation (LAD), T-wave abnormality (TAb), atrial fibrillation (AF), and complete right bundle branch block (CRBBB). After filtering to this shared label space the clients contain 4108 (CPSC), 5489 (Georgia), and 20517 (PTB-XL) records.

Figure 2 shows the within-client prevalence of the five labels. The distribution is markedly heterogeneous: PTB-XL is dominated by sinus rhythm (∼88%), CPSC by CRBBB (∼48%) and AF (∼33%), and Georgia by T-wave abnormality (∼42%). Combined with PTB-XL holding roughly two thirds of all records, this creates both quantity and label-distribution skew across clients, which is the setting label-aware aggregation is designed for.

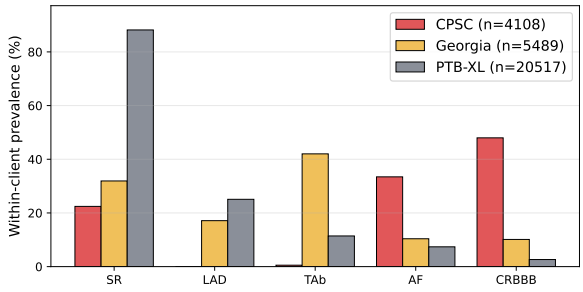

Figure 2. Within-client prevalence of the five shared diagnoses. The label distribution is strongly non-IID across clients: PTB-XL is dominated by sinus rhythm, CPSC by CRBBB and AF, and Georgia by T-wave abnormality, while PTB-XL also holds the large majority of all records.

2.1. Training Framework

All experiments use *CTNLite*, a lightweight port of the wide-and-deep convolution-transformer that won the original CinC 2020 Challenge [6]. *CTNLite* keeps the convolutional encoder and a transformer encoder with mean pooling over time followed by fully connected layers, but drops the original model's fusion of demographic and handcrafted waveform features, so that it trains directly on raw waveforms. All models are trained with Adam (learning rate 10^{-3}) under a binary cross-entropy objective over the five target labels.

Padding ablation: CPSC recordings are highly variable in length (up to ∼118 s), whereas Georgia and PTB-XL recordings are almost uniformly 10 s. We therefore compare a fixed-length crop/pad pipeline (`fixed+wrap`) against a `batch_dynamic` scheme that pads each mini-batch only to its own maximum length and masks padded positions in the transformer, with two fill strategies (`wrap` and `zero`). This yields three padding configurations; the results below use `fixed+wrap` unless noted, and the per-client pattern is consistent across all three.

Federated learning algorithms: We compare five training modes. **Central** pools all client data into a single model, an upper-reference not subject to communication constraints. **FedAvg** aggregates client updates with weights proportional to local training-set size [2]. **FedProx** adds an ℓ_2 proximal term ($\mu = 0.01$) to each client's local objective, penalizing drift from the global model, and otherwise aggregates as FedAvg [3]. **FedLA** replaces the aggregation weights with label-distribution-derived weights: for each label, each client's share of that label's total count across clients is computed, summed over labels, and normalized across clients [4]. **FedProx+LA** combines the FedProx local proximal term with FedLA's aggregation weights. All federated modes use one local epoch per communication round.

Local-only baseline: To quantify what federation actually helps each site, we additionally train a model on a single client's data alone and evaluate it on that same client's hold-out data, for every client, fold, and padding configuration. These models use the identical architecture, optimizer, and schedule as the federated runs, and are constrained to the same five-label output space derived from the pooled data, so that their scores remain directly comparable to the federated per-client numbers.

Experimental protocol: Each of the three padding configurations $\times$ five training modes is evaluated with 5-fold cross-validation on the full client data, 30 training epochs/communication rounds per fold. Within each fold, 20% of a client's records are set aside as the hold-out first, and the remaining data is further split into a 85% training and 15% validation portion, which represents 68% and 12% of all data respectively. The validation split is not used for model selection or threshold tuning in any reported result, so every number reflects the model at the final epoch / round rather than a validation-selected checkpoint. To avoid data leakage, all federated modes split each client's data independently, so a client's hold-out never overlaps with its own training or validation data. Central instead pools all clients before splitting, so its partition is comparable in expectation but not identical to the federated one. We report AUROC (macro, threshold-independent) as mean $\pm$ standard deviation across the five folds, computed both on the pooled hold-out and on each client's own hold-out.

3. Results

Padding ablation: Table 1 reports the pooled (Global) AUROC for all five training modes under the three padding configurations described in Methods. No single padding configuration wins outright: `batch_dynamic+wrap` is best for FedAvg and FedProx, and `batch_dynamic + zero` is best for Central, FedLA, and FedProx+LA. However, `fixed+wrap` stays close behind the winning configuration for three of the five modes – FedAvg, FedProx, and FedLA – trailing the best alternative by no more than 0.005 AUROC in each case. We use `fixed+wrap` as the paper's primary configuration because it remains competitive for most of the modes we study, not because it is uniformly the strongest padding strategy.

Table 1. Global (pooled) AUROC (5-fold mean $\pm$ std) by padding configuration and training mode. Bold marks the best configuration per mode.

Mode	fixed+wrap	batch_dyn.+wrap	batch_dyn.+zero
Central	0.954±0.002	0.955±0.002	**0.960±0.003**
FedAvg	0.929±0.002	**0.930±0.003**	0.929±0.003
FedProx	0.928±0.003	**0.930±0.003**	0.928±0.004
FedLA	0.931±0.003	0.933±0.002	**0.933±0.001**
FedProx+LA	0.923±0.006	0.923±0.003	**0.929±0.004**

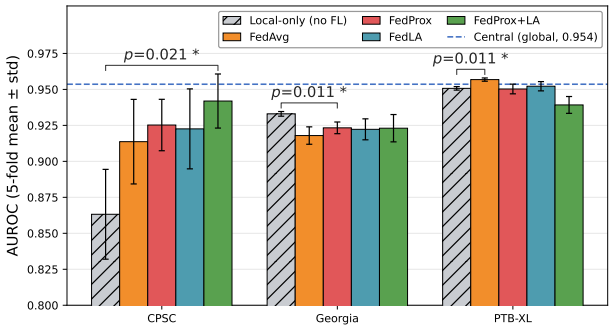

Figure 3. Per-client AUROC (5-fold mean $\pm$ std, `fixed+wrap`). Hatched bars are local-only training, colored bars are the four federated modes, and the dashed line is central pooled AUROC. Brackets give the paired t-test between local-only training and that client's best federated mode, Holm-corrected across the three comparisons. Georgia's significant difference is a decrease.

Per-client performance: Figure 3 and Table 2 report the 5-fold AUROC for each client under `fixed+wrap`, for the four federated modes and for the local-only baseline, with central pooled training as an upper reference. On the pooled hold-out the four federated modes lie within 0.008 AUROC of each other, so no aggregation rule stands out there. However, the per-client view shows a different story. On PTB-XL, the largest client, FedAvg is the strongest federated mode, which is consistent with a sample-size weighting that matches the client holding roughly two thirds of all records. On Georgia, the mid-sized client, the four modes lie within 0.006 of each other and every one of them falls below the local-only baseline.

Table 2. Per-client AUROC (5-fold mean $\pm$ std, `fixed+wrap`): local-only training versus the best-performing federated mode for that client. Δ is best-federated minus local-only; p is a paired t-test across the five folds, Holm-corrected over the three per-client comparisons.

Client	Local-only	Best federated	Δ	p
CPSC	0.863±0.031	0.942±0.019 (FedProx+LA)	+0.079	0.021
Georgia	0.933±0.002	0.923±0.004 (FedProx)	−0.010	0.011
PTB-XL	0.951±0.001	0.957±0.001 (FedAvg)	+0.006	0.011

On CPSC, the smallest and most label-skewed client, the separation is the largest: FedProx+LA is the strongest mode, followed by FedProx and, FedLA while FedAvg is the weakest. The three clients therefore benefit very unequally from federation, and the sign of Δ flips between them. CPSC gains 0.079 AUROC over training alone, PTB-XL gains 0.006, and Georgia loses 0.010; all three differences are significant under `fixed+wrap`. How much CPSC gains depends on which aggregation rule is used, and not only on federating at all: FedProx+LA re-

covers $+0.079$ AUROC where plain FedAvg recovers only $+0.050$.

4. Discussion

This work shows that reporting only the overall pooled performance gain does not tell the whole story, especially for realistic applications such as hospital settings. The per-client breakdown is essential to interpret these results, since the four aggregation rules look indistinguishable on the pooled hold-out alone, while on CPSC and PTB-XL they move in opposite directions. Each aggregation method has its own strength. FedAvg and FedProx weight clients by their number of training samples, so PTB-XL – roughly two thirds of all records – benefits from an aggregate that reflects its own size and label prior. FedLA and FedProx+LA instead weight clients by their share of each label, which up-weights CPSC despite its small size because it holds a disproportionate share of the CRBBB and AF examples. The resulting takeaway is that no aggregation method is best everywhere: which rule helps depends on how a client's data distribution relates to the federation as a whole, and for Georgia federation offers no benefit under any rule we tested. FedProx+LA even scores below PTB-XL's own local-only baseline (0.939 versus 0.951), so a method that helps one client can actively cost another. Still, for two of the three clients some form of federation improved on training alone, which lets us conclude that there is a genuine advantage to federated training: hospitals can benefit from each other's data without ever sharing it, provided the aggregation rule is matched to the setting.
Limitations: Every federated mode remains below pooled central training (0.954 versus 0.923–0.931), so the privacy constraint still carries an accuracy cost that aggregation choice does not recover. CPSC's gain is also the least robust result we report: it reaches significance only under `fixed+wrap`, because its local-only baseline is both stronger and about twice as variable under batch-dynamic padding. Its macro AUROC additionally rests on fewer effective labels – one of five has no positive hold-out examples and a second only a handful – which explains its wider error bars. The federation comprises only three real institutional clients, too few to characterize how the effect scales with client count. The label space is restricted to five diagnoses, and FedProx's μ and the local epoch count were fixed rather than tuned per client or mode.

5. Conclusion

We benchmarked central, FedAvg, FedProx, FedLA, and FedProx+LA on three heterogeneous PhysioNet/CinC 2020 sources under 5-fold cross-validation, together with a local-only per-client baseline. The pooled metric alone would suggest the aggregation rule barely matters; the per-client view shows the opposite, with each rule favoring a different client depending on its size and label distribution. Federated learning is therefore worth adopting here, but not as a fixed recipe: each site should check against its own local-only model whether it gains at all, and the aggregation rule should be matched to the label distribution of the federation rather than chosen once for all clients. In this sense the aggregation rule is a fairness lever, not a uniform accuracy gain. Since label-aware aggregation was originally proposed for vehicular object detection, our results also show that the idea transfers to multi-label clinical ECG classification. We see the full 27-class label space, more participating sites, and rarer diagnoses as the clearest extension, since label-aware weighting may matter even more there.

Funding

This work was financed by the HEAD-Genuit-Stiftung and CVDLink.

References

[1] Di Cesare M, Perel P, Taylor S, Kabudula C, Bixby H, Gaziano TA, McGhie DV, Mwangi J, Pervan B, Narula J, Pineiro D, Pinto FJ. The Heart of the World ;19(1):11. ISSN 2211-8179.

[2] McMahan B, Moore E, Ramage D, Hampson S, Aguera B. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Singh A, Zhu J (eds.), Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, volume 54 of Proceedings of Machine Learning Research. PMLR; 1273–1282.

[3] Li T, Sahu AK, Zaheer M, Sanjabi M, Talwalkar A, Smith V. Federated optimization in heterogeneous networks ;2:429–450.

[4] Khalil A, Dege T, Golchin P, Olshevskyi R, Anta AF, Meuser T. Federated Learning with Heterogeneous Data Handling for Robust Vehicular Object Detection.

[5] Perez Alday EA, Gu A, Shah A, Liu C, Sharma A, Seyedi S, Bahrami Rad A, Reyna M, Clifford G. Classification of 12-lead ECGs: The PhysioNet/Computing in Cardiology Challenge 2020.

[6] Natarajan A, Chang Y, Mariani S, Rahman A, Boverman G, Vij S, Rubin J. A Wide and Deep Transformer Neural Network for 12-Lead ECG Classification; .

Address for correspondence:

Tizian Dege
Merckstraße 25
dege@kismed.tu-darmstadt.de