Comparing Centralized and Federated Learning for ECG-Based Chagas Disease Detection

Thierry Bossy[2][0000−0001−8442−4705], Andrea Loddo[1][0000−0002−6571−3816], Mirko Marras[1][0000−0003−1989−6057], Vittoria Pala[1], and Alessandra Perniciano[1][0009−0003−8956−5058] ✉

[1] Department of Mathematics and Computer Science, University of Cagliari, Cagliari, Italy
[2] Tune Insight SA, Switzerland
✉alessandra.pernician@unica.it

Abstract. Chagas disease is a neglected tropical infection caused by *Trypanosoma cruzi*, affecting approximately eight million people. Although serology is the diagnostic standard, its cost limits population-wide screening. The inexpensive 12-lead electrocardiogram (ECG) offers a promising alternative, but existing deep learning approaches typically rely on centralized data pooling, which may conflict with privacy regulations. We present a controlled benchmark of centralized and federated learning for ECG-based Chagas disease detection, progressively partitioning patient-level data across clients while keeping the architecture, preprocessing, loss, optimizer, and basic hyperparameters fixed while using predefined centralized and federated training schedules. On CODE-15%, federated learning closely matches centralized performance with two clients (AUPRC = 0.150, 7.6× prevalence lift), but degrades with increasing fragmentation: at 100 clients, AUROC falls from 0.836 to 0.664 and the precision–recall lift to 1.7×. The associated increase in false positives depends on the selected operating threshold. The results are consistent with reduced exposure to positive examples contributing to degradation, although the present design does not isolate this mechanism from other effects of increased client fragmentation. **Source code**: `https://github.com/tail-unica/federated-learning-chagas`.

Keywords: Federated learning · Chagas disease · Electrocardiography.

1 Introduction

Chagas disease is a tropical parasitic disease caused by *Trypanosoma cruzi*, affecting around eight million people, predominantly in Latin America, and causing over ten thousand deaths annually [22]. Up to 45% of individuals with chronic infections develop Chagas cardiomyopathy 10–30 years after the initial exposure, often progressing to heart failure. During the acute phase, rapid parasite proliferation yields high parasitemia, making diagnosis straightforward via peripheral blood smear, while specific abnormalities can already be observed on the electrocardiogram (ECG) [11]. However, the disease often remains asymptomatic

for decades, leaving 70% of those infected unaware of their condition [4]. When symptoms eventually emerge in the chronic phase, the parasite hides within cardiac and digestive muscle cells to evade the immune system, making it undetectable in direct blood smears. Confirming infection at this stage requires serological testing, which depends on laboratory infrastructures that are scarce where the disease burden is highest [21]. Since interventions are most effective before irreversible cardiac damage occurs, early detection is critical, creating a need for an inexpensive tool to identify candidates for confirmatory testing.

The 12-lead ECG is a prime candidate, since chronic involvement of the myocardium and of the conduction system leaves characteristic abnormalities and deep networks detect infected individuals from the ECG alone [7,14]. The primary limitation is data locality rather than accuracy: existing approaches train centrally, whereas the records sit in institutions that regulation prevents from pooling. Federated learning enables these institutions to train a shared model while records stay local [10]. However, to the best of our knowledge, federated learning has not yet been applied to Chagas disease from ECG signals. We systematically benchmark the cost of decentralized training against a centralized upper bound to analyze the performance gap as data fragments. Holding all non-training components fixed, our comparison extends from two to one hundred clients, reaching a regime where positive examples per client become scarce. To isolate the effect of client fragmentation, we adopt a controlled setting using FedAvg and standardized partitioning. Rather than reproducing full institutional heterogeneity—such as non-IID prevalence, feature shifts, unequal client sizes, or cross-site acquisition effects—this design establishes a clear baseline for understanding how fragmentation affects federated performance. We address two questions: the performance gap between a centralized model and its federated counterpart under identical conditions (**RQ1**), and how this gap scales with the number of clients (**RQ2**).

2 Related Work

Recurrent networks were the earliest deep models applied to the ECG [18], before convolutional layers took over local feature extraction; Zihlmann et al. [24] combined the two families and showed the hybrid to outperform a purely convolutional counterpart, and at full-record scale a deep residual network recognises several abnormalities from recordings drawn from the same Brazilian telehealth network that supplies the CODE 15% dataset we use also in our study in this paper [14]. Work on Chagas disease is recent and, without exception, centralized: a deep network screens for it from the ECG with a sharp drop over independent cohorts [7]. More recently, the task became a community benchmark [13].

Since these ECG records cannot be pooled across institutions, federated learning offers a natural remedy, training a shared model while the data stay local. Its standard algorithm, FedAvg [10], aggregates client updates in proportion to local data; under non-IID data, the local models drift apart. FedProx [9] and SCAFFOLD [8] correct this limitation through a proximal term and control

variates, respectively, both of which matter only to the extent that the partition is heterogeneous. Exchanging updates does not, by itself, protect the clients; hence, secure aggregation [2] and differential privacy [1] are often used. Class imbalance poses a further challenge: severe local scarcity of positive cases degrades model learning and accelerates inter-client model drift even under an IID partition [20]. The study closest to ours proposes to train arrhythmia classifiers on ECGs held by six separate sources and vary the client count from two to ten [6]; our study in this paper differs in four main perspectives: its target disease, in using an end-to-end network on the raw signal, in extending the client range to one hundred, and in the single-source data design.

3 Methodology

In our study, the centralized and federated configurations share the same underlying data, preprocessing, architecture, loss function, and evaluation protocol, while the training procedure differs between paradigms. The centralized model is used as an internal reference for assessing the effect of distributed training under otherwise matched experimental conditions, rather than as a state-of-the-art centralized benchmark.

3.1 Dataset Preprocessing

We evaluated our approach using CODE-15% from the 2025 PhysioNet/CinC Challenge [12], a stratified 15% sample of the CODE cohort of 12-lead ECGs sampled at 400 Hz by the Telehealth Network of Minas Gerais [14,15]. Among the challenge datasets, CODE-15% was preferred over SaMi-Trop [16] and PTB-XL [19] because it is the only single-source dataset containing both target classes. This design provides experimental control by preventing differences in acquisition hardware, filtering, sampling, and lead configuration from becoming additional explanatory factors when assessing client fragmentation, thereby reducing potential shortcut-learning effects [5,17]. Chagas labels are linked to individual ECG examinations through the examination identifier and are based on self-reported diagnoses. Because multiple ECG recordings may be available for the same patient, all dataset splitting and client partitioning are performed at the patient level, ensuring that recordings from a given patient remain within a single partition. Moreover, CODE-15% aggregates records from telehealth centers across hundreds of municipalities, enabling patient-level partitioning into disjoint nodes that is compatible with a federated setting. This does not capture the full heterogeneity of real institutional federations, which is left for future work.

For preprocessing, all records were converted to WFDB format. Unlabelled or excessively short recordings were discarded, and the remaining signals were truncated to 2,800 samples. A zero-phase cascade [3] was applied independently to each lead, comprising a 60 Hz notch and third-order Butterworth high- and

Table 1. Data splits (left) and federated partition statistics across client configurations (right). Splits are patient-level and stratified by class; the test set is shared across all configuration while the federated pool comprises the union of the centralized training and validation sets. Ranges over clients are in brackets.

Dataset splits			
Partition	Records	Pos.	Patients
Working set	342,655	6,559	233,178
Centralized train (70%)	239,744	4,521	—
Centralized val (10%)	34,189	682	—
Shared test (20%)	68,722	1,356	46,637
Federated pool	273,933	5,203	186,541

Federated statistics			
N	Train rec.	Train pos.	Val. pos.
2	119,819	2,283 [2,220–2,346]	∼318
10	23,963	457 [435–482]	∼63
20	11,995	228 [208–245]	∼33
50	4,792	90 [71–103]	∼15
100	2,398	45 [33–61]	∼7

low-pass filters at 5 and 60 Hz. The notch was retained because the Butterworth roll-off only gradually attenuates frequencies near its cut-off, leaving residual 60 Hz power-line interference. All filtering stages were applied forward and backward to avoid timing and morphology distortions, which is important for the time-frequency spectrogram front-end. Finally, splitting was performed at the patient level to prevent data leakage (Table 1). The data were split at the patient level into 70% training, 10% validation, and 20% shared test sets. The test set was kept identical across paradigms. The federated training pool corresponded exactly to the union of the centralized training and validation sets and was subsequently partitioned across clients at the patient level; within each client, 87.5% of patients were assigned to training and 12.5% to validation. Thus, up to patient-level rounding, the aggregated federated training and validation sets correspond to those of the centralized setting.

3.2 Dataset Federated Splitting

To simulate an Independent and Identically Distributed (IID) federated setting across $N \in \{2, 10, 20, 50, 100\}$ clients, we partitioned the data by assigning positive and negative patients separately in a round-robin order. This approach ensures that each client preserves the global class ratio while using approximately balanced shard sizes as a controlled experimental assumption (coefficient of variation below 0.02 even at 100 clients). This assumption was adopted to isolate the effect of client count from the additional variability introduced by strongly unequal local dataset sizes.

To replicate the 70/10 split used in the centralized setting for training and validation, each client data is further split into 87.5% of patients to training and 12.5% to validation. This guarantees that the aggregate federated data match the centralized one, thereby keeping the data membership fixed while comparing the centralized and federated training protocols. The choice of the number of clients follows a geometric progression whose upper bound shifts the environment into a low-data regime where each client holds only a few tens of positive records. Being IID by design, this setup varies the statistical fragmentation of the positive class without introducing distributional heterogeneity; it is therefore a controlled baseline rather than a representation of typical institutional federation.

3.3 Model Architecture Selection

The model architecture was selected based on preliminary central evaluations comparing seven candidates, encompassing recurrent, convolutional-recurrent, and time-frequency representations. All models were trained centrally under a shared configuration on a balanced subset of 10,000 records from CODE-15%. The selected model, a hybrid recurrent convolutional network, achieved the highest AUROC. Its front-end computes per lead a short-time Fourier transform (Hann window 64, hop 32, FFT length 256), takes the magnitude, applies a compressive $\log(1 + 128\,|\cdot|)$ transformation and re-bins the frequency axis onto 64 logarithmically spaced bins. Two 3×3 convolutional layers with 32 and 64 channels and 2×2 max-pooling extract local time–frequency patterns, temporal mean-pooling reduces the feature map, and two stacked unidirectional gated recurrent units of 32 units aggregate it before a sigmoid output, for 37,761 parameters. This compact size was selected to keep per-round communication overhead low and limit the memorisation of client-specific artifacts.

3.4 Model Training and Optimisation

Both paradigms are trained using their respective predefined training schedules: 100 central epochs versus 20 federated rounds of 5 local epochs, and stochastic gradient descent with momentum (learning rate 10^{-3}, momentum 0.9, batch size 128, no weight decay). These schedules define the training protocols for the two settings and are not intended to represent equivalent optimizer trajectories.

To handle class imbalance without altering the local distributions, unlike resampling, both settings share a global weighted binary cross-entropy loss with a positive weight of ~ 52 (the negative-to-positive ratio of the training pool). This keeps the optimization objective identical across paradigms and remains compatible with secure aggregation. Each round uses standard FedAvg [10] weighted by local dataset size. It should be noted that aggregation is performed in *cleartext*, meaning that no privacy mechanisms are applied. A single random seed was used to fix shuffling, partitioning, and initialization, and each configuration was trained once. Consequently, the reported results do not capture variability across independent training runs or alternative random partitions.

The experimental design was intentionally controlled: FedAvg, approximately balanced client sizes, and a unified data source were kept fixed while varying the number of clients. This isolates client fragmentation but does not capture non-IID distributions, unequal client sizes, or cross-site acquisition heterogeneity typical of institutional federations.

3.5 Model Evaluation

The performance is evaluated using Area Under the Receiver Operating Characteristic Curve (AUROC), complemented by the Area Under the Precision-Recall Curve (AUPRC), which offers better informativeness under severe class

Table 2. Test-set performance and paired centralized-minus-federated gap. AUPRC is reported as a lift over the test prevalence baseline (0.0197). Operating points (recall, specificity, and threshold) are selected by maximizing Youden's J statistic on validation data. Confidence intervals (95%) represent stratified paired-bootstrap estimates on score difference; DeLong's test gives $p = 0.003$ at $N = 2$ and $p < 0.001$ for all others.

	Absolute performance					*Paired gap vs. centralized*	
Configuration	AUROC	AUPRC (lift)	Recall	Spec.	Thr.	ΔAUROC [95% CI]	ΔAUPRC [95% CI]
Centralized	0.8355	0.1502 (7.6×)	0.739	0.753	0.627	—	—
Federated $N = 2$	0.8252	0.1496 (7.6×)	0.777	0.692	0.464	0.0103 [+0.0036, +0.0172]	0.0003 [−0.0116, +0.0124]
Federated $N = 10$	0.8157	0.1208 (6.1×)	0.768	0.689	0.422	0.0198 [+0.0131, +0.0264]	0.0296 [+0.0183, +0.0416]
Federated $N = 20$	0.7702	0.0859 (4.4×)	0.684	0.704	0.514	0.0652 [+0.0558, +0.0750]	0.0646 [+0.0525, +0.0775]
Federated $N = 50$	0.6852	0.0374 (1.9×)	0.646	0.620	0.542	0.1503 [+0.1372, +0.1641]	0.1135 [+0.0993, +0.1292]
Federated $N = 100$	0.6639	0.0344 (1.7×)	0.782	0.447	0.506	0.1717 [+0.1571, +0.1862]	0.1166 [+0.1022, +0.1324]

imbalance, and the Expected Calibration Error (ECE). AUPRC values are interpreted against the baseline test prevalence (0.0197). The decision threshold is determined by maximizing Youden's J statistic [23] on validation data and applied once to the test set. Empirical thresholding overcomes the over-confidence induced by the weighted loss, as miscalibration shifts score values without altering their ordering. "Statistical significance is evaluated using a stratified paired bootstrap on score differences, cross-checked with DeLong's test for AUROC. These procedures quantify uncertainty in the test-set comparison but do not capture variability across independent training runs or random client partitions. The held-out test set of 68,722 records is evaluated only once, after restoring the checkpoint with the highest validation AUPRC, to ensure a fair comparison.

4 Experimental Results

In this section, we analyze the experimental results to address our two primary research questions regarding the quantitative and qualitative gap between centralized and federated paradigms for ECG-based Chagas disease detection.

4.1 RQ1: Centralized–Federated Performance Gap

In the first analysis, we investigated the performance gap between a centralized model and its federated counterpart under identical experimental conditions. Table 2 summarizes the results. AUROC declines monotonically with the number of clients, from 0.836 centrally to 0.825 with two clients and 0.664 with 100 clients. AUPRC follows the same ordering and decreases from 7.6× to 1.7× the centralized prevalence baseline. We also report recall, specificity, and the threshold selected by Youden's J. Recall is higher for the two least fragmented configurations ($N = 2, 10$) than for the centralized model, but this reflects threshold selection: the federated thresholds (0.464 and 0.422) are lower than the centralized threshold (0.627), increasing positive predictions at the cost of specificity and more false positives. In Figure 1, models remain relatively close in ROC space but separate more clearly in precision–recall space. We therefore focus primarily on threshold-free metrics.

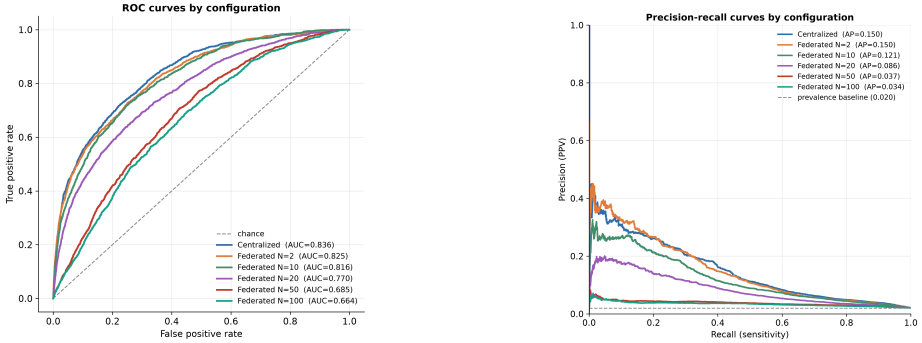

Fig. 1. ROC (left) and precision–recall (right) curves on the shared test set. The dashed line denotes the test-set prevalence baseline (0.0197).

Paired statistical analysis confirms that the AUROC differences exceed sampling variability. The AUROC gap is significant already at $N = 2$ and increases at every subsequent fragmentation level. Paired evaluation is essential: although the marginal 95% CIs of the centralized baseline ($[0.8244, 0.8461]$) and two-client model ($[0.8140, 0.8364]$) overlap, their paired difference excludes zero ($[+0.0036, +0.0172]$). Conversely, at $N = 2$, the AUPRC gap is not significant ($[-0.0116, +0.0124]$), indicating that the two-client federation matches the centralized model on the primary screening metric despite a measurable reduction in ranking performance. From $N = 10$ onward, the AUPRC gap is also significant. All differences remain significant after Bonferroni correction for the five pairwise comparisons.

Finally, the weighted loss might suggest that AUROC degradation is related to miscalibration, since a positive-class weight near 52 inflates positive scores. ECE distinguishes these effects: discrimination declines monotonically with fragmentation, whereas ECE does not, measuring 0.35 centrally, 0.29 at $N = 2$ and $N = 10$, and increasing only under heavier fragmentation (0.41 at $N = 20$, 0.49 at $N = 50$ and $N = 100$). Across the range where the discrimination gap first emerges, ECE is not monotonically related to the performance gap, suggesting that the observed degradation cannot be explained solely by calibration differences. Nevertheless, the high absolute ECE values indicate substantial miscalibration across configurations, making calibration an important limitation of the present models.

4.2 RQ2: Scaling of the Centralized–Federated Performance Gap

In the second analysis, we examined how the centralized–federated performance gap changes with the number of clients. We additionally report inter-client dispersion in local validation performance as a descriptive analysis. The analysis is descriptive and does not directly measure parameter- or update-space divergence. Results from Table 2 show that the paired performance gap increases unevenly. The AUROC gap grows from 0.010 ($N = 2$) to 0.020 ($N = 10$), then

rises sharply to 0.065 ($N = 20$), 0.150 ($N = 50$), and 0.172 ($N = 100$). AUPRC follows a similar trend, increasing from a non-significant gap at $N = 2$ to 0.117 at $N = 100$. On a logarithmic scale, degradation accelerates between $N = 10$ and $N = 20$ and plateaus beyond $N = 50$. This corresponds to a sharp reduction in precision–recall lift (7.6× to 1.7×), showing that an AUROC of 0.685 ($N = 50$) can conceal a substantial practical cost.

We additionally examined two descriptive aspects that accompany this degradation. Under statistical fragmentation, near-equal-sized shards reduce positive examples per client from about 2,283 at $N = 2$ to 45 at $N = 100$. Reduced positive exposure is therefore a plausible contributor, although the present design does not isolate it from other consequences of increasing client count. As shown in Figure 2, inter-client AUROC standard deviation increases from 0.019 ($N = 2$) to 0.114 ($N = 100$), indicating greater dispersion in local predictive performance rather than a uniform decline (at $N = 50$, performance ranges from 0.51 to 0.83). This metric measures performance dispersion and should not be interpreted as direct parameter- or update-space divergence. Client shard sizes were approximately balanced by construction ($CV < 0.02$), so unequal client quantity was not a source of heterogeneity in the present experiment.

These observations are associational: positive-class fragmentation and inter-client performance dispersion increase together with N, establishing co-variation rather than causality. The shard-size analysis likewise provides no evidence for a size-heterogeneity explanation under the present partitioning scheme. We therefore interpret reduced positive exposure as a plausible contributor without making a causal attribution. Operationally, at $N = 100$, the model detects 1,060 of 1,356 positive cases, compared with 1,002 centrally, but produces 37,226 false positives versus 16,632. Because these operating points are independently selected using Youden's J, the increase in false positives is conditional on the selected thresholds and should not be interpreted as a threshold-independent estimate of screening burden.

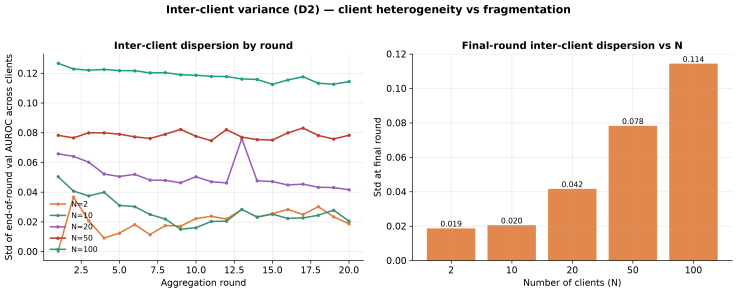

Fig. 2. Descriptive dispersion of local validation AUROC. At high client counts, local validation sets contain few positive examples, so the statistic is increasingly affected by sampling variability and should not be interpreted as parameter or update divergence.

5 Conclusions and Future Work

We benchmarked centralized and federated learning for ECG-based Chagas disease detection under controlled IID conditions. Federated performance remained close to the centralized reference with two clients, but degraded substantially as the data were increasingly fragmented, with AUROC decreasing from 0.836 centrally to 0.664 at 100 clients.

The decrease in positive examples per client is consistent with reduced positive exposure contributing to this degradation, but the present design does not allow this mechanism to be causally isolated. Similarly, inter-client validation AUROC dispersion should be interpreted as predictive-performance variability rather than parameter or update divergence.

The controlled use of a unified data source, approximately balanced IID partitions, and fixed FedAvg isolates client fragmentation but limits generalizability to heterogeneous institutional settings. Future work should investigate non-IID data, unequal client sizes, alternative aggregation strategies, and direct measures of model divergence, while also addressing multi-seed variability and patient-level statistical dependence. Lastly, assessing the clinical sensitivity of our preprocessing choices, and evaluating alternative signal-filtering strategies in particular, remains an open area for future inquiry.

Disclosure of Interests. Thierry Bossy is an employee of Tune Insight. Vittoria Pala conducted an unpaid internship at Tune Insight during the development of this work. The remaining authors declare no competing interests.

References

1. Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L.: Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS). pp. 308–318 (2016). https://doi.org/10.1145/2976749.2978318
2. 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: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS). pp. 1175–1191 (2017). https://doi.org/10.1145/3133956.3133982
3. Donida Labati, R., Sassi, R., Scotti, F.: ECG biometric recognition: Permanence analysis of QRS signals for 24 hours continuous authentication. In: 2013 IEEE International Workshop on Information Forensics and Security (WIFS). pp. 31–36 (2013). https://doi.org/10.1109/WIFS.2013.6707790
4. GBD 2023 Chagas Disease and RAISE Study Collaborators: Global, regional, and national burden of Chagas disease, 1990–2023: a systematic analysis for the Global Burden of Disease Study 2023. The Lancet Infectious Diseases **26**(3), 284–301 (2026). https://doi.org/10.1016/S1473-3099(25)00562-6
5. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11), 665–673 (2020). https://doi.org/10.1038/s42256-020-00257-z

6. Gutierrez, D.M.J., Hassan, H.M., Landi, L., Vitaletti, A., Chatzigiannakis, I.: Application of federated learning techniques for arrhythmia classification using 12-lead ecg signals (2022), https://arxiv.org/abs/2208.10993
7. Jidling, C., Gedon, D., Schön, T.B., Di Lorenzo Oliveira, C., Cardoso, C.S., Ferreira, A.M., Giatti, L., Barreto, S.M., Sabino, E.C., Ribeiro, A.L.P., Ribeiro, A.H.: Screening for Chagas disease from the electrocardiogram using a deep neural network. PLOS Neglected Tropical Diseases **17**(7), e0011118 (2023). https://doi.org/10.1371/journal.pntd.0011118
8. Karimireddy, S.P., Kale, S., Mohri, M., Reddi, S.J., Stich, S.U., Suresh, A.T.: SCAFFOLD: Stochastic controlled averaging for federated learning. In: Proceedings of the 37th International Conference on Machine Learning (ICML). PMLR, vol. 119, pp. 5132–5143 (2020)
9. Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V.: Federated optimization in heterogeneous networks. In: Proceedings of Machine Learning and Systems (MLSys). vol. 2, pp. 429–450 (2020)
10. McMahan, B., Moore, E., Ramage, D., Hampson, S., Agüera y Arcas, B.: Communication-efficient learning of deep networks from decentralized data. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS). Proceedings of Machine Learning Research, vol. 54, pp. 1273–1282 (2017)
11. Nunes, M.C.P., Beaton, A., Acquatella, H., Bern, C., Bolger, A.F., Echeverría, L.E., Durante-Mangoni, E., Figueiredo, H.L., Gilman, R.H., Santos, R.D., et al.: Chagas cardiomyopathy: an update of current concepts. Critical Reviews in Clinical Laboratory Sciences **55**(7), 435–468 (2018)
12. PhysioNet/Computing in Cardiology Challenge: Detection of Chagas disease from the ECG: The George B. Moody PhysioNet Challenge 2025 (2025), https://moody-challenge.physionet.org/2025, accessed: July 2026
13. Reyna, M.A., Koscova, Z., Pavlus, J., Weigle, J., Saghafi, S., Gomes, P., Elola, A., Hassannia, M.S., Campbell, K., Bahrami Rad, A., Ribeiro, A.H., Ribeiro, A.L.P., Sameni, R., Clifford, G.D.: Detection of Chagas disease from the ECG: The George B. Moody PhysioNet Challenge 2025. In: Computing in Cardiology. vol. 52, pp. 1–4 (2025)
14. Ribeiro, A.H., Horta Ribeiro, M., Paixão, G.M.M., Oliveira, D.M., Gomes, P.R., Canazart, J.A., Ferreira, M.P.S., Andersson, C.R., Macfarlane, P.W., Meira Jr., W., Schön, T.B., Ribeiro, A.L.P.: Automatic diagnosis of the 12-lead ECG using a deep neural network. Nature Communications **11**(1), 1760 (2020). https://doi.org/10.1038/s41467-020-15432-4
15. Ribeiro, A.H., Paixão, G.M.M., Lima, E.M., Horta Ribeiro, M., Pinto Filho, M.M., Gomes, P.R., Oliveira, D.M., Meira Jr., W., Schön, T.B., Ribeiro, A.L.P.: CODE-15%: a large scale annotated dataset of 12-lead ECGs. Zenodo (2021), https://doi.org/10.5281/zenodo.4916206
16. Ribeiro, A.L.P., Ribeiro, A.H., Paixão, G.M.M., Lima, E.M., Horta Ribeiro, M., Pinto Filho, M.M., Gomes, P.R., Oliveira, D.M., Meira Jr., W., Schön, T.B., Sabino, E.C.: SaMi-Trop: 12-lead ECG traces with annotations. Zenodo (2021), https://doi.org/10.5281/zenodo.4905618
17. Strodthoff, N., Wagner, P., Schaeffter, T., Samek, W.: Deep learning for ecg analysis: Benchmarks and insights from ptb-xl. IEEE Journal of Biomedical and Health Informatics **25**(5), 1519–1528 (2020)
18. Übeyli, E.D.: Combining recurrent neural networks with eigenvector methods for classification of ECG beats. Digital Signal Processing **19**(2),

320–329 (2009). https://doi.org/10.1016/j.dsp.2008.09.002, https://www.sciencedirect.com/science/article/pii/S1051200408001516
19. Wagner, P., Strodthoff, N., Bousseljot, R.D., Kreiseler, D., Lunze, F.I., Samek, W., Schaeffter, T.: PTB-XL, a large publicly available electrocardiography dataset. Scientific Data **7**(1), 154 (2020). https://doi.org/10.1038/s41597-020-0495-6
20. Wang, L., Xu, S., Wang, X., Zhu, Q.: Addressing class imbalance in federated learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 10165–10173 (2021)
21. World Health Organization: Integrating neglected tropical diseases into primary health care: a manual for health workers. World Health Organization, Geneva (2022), https://iris.who.int/handle/10665/351421
22. World Health Organization: Chagas disease (also known as American trypanosomiasis). WHO Fact Sheet (2026), https://www.who.int/news-room/fact-sheets/detail/chagas-disease-(american-trypanosomiasis), accessed July 2026
23. Youden, W.J.: Index for rating diagnostic tests. Cancer **3**(1), 32–35 (1950). https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3
24. Zihlmann, M., Perekrestenko, D., Tschannen, M.: Convolutional recurrent neural networks for electrocardiogram classification. In: 2017 Computing in Cardiology (CinC). pp. 1–4. IEEE (2017). https://doi.org/10.22489/CinC.2017.070-060