A Broader Impact

Considering that this research exclusively involves the repurposing of existing open-source databases, the associated risks are limited. However, it is important to acknowledge that all datasets utilized in this study may be influenced by biases inherent in the original data collection processes, such as those related to gender, age, or race. Unfortunately, identifying the sources of potential biases is challenging because the data have been appropriately pseudonymized. Moreover, records such as electrocardiograms and echocardiograms cannot be easily linked to specific demographic attributes such as age, ethnicity, or gender by non-medical experts. Nonetheless, our work discloses certain metadata of the datasets, including geographical origin, gender distribution, and age distribution. This exposure may aid in identifying underlying geographical biases, which are anticipated in real-world federated learning scenarios.

While prioritizing simplicity and utility, the current benchmark does not include privacy metrics. Nevertheless, privacy remains critically important in the cardiovascular disease domain, and we strongly encourage the research community to address these considerations. Thanks to the modularity of FedCVD, we can add privacy components easily. Therefore, we anticipate that FedCVD will address privacy concerns related to federated learning within the cardiovascular disease domain in the future.

B Datasets repository and Maintenance plane

B.1 Dataset repository.

The code is now available at `https://anonymous.4open.science/r/ZYNTMBB-8848`. Considering licenses, users need to download the data manually through the original dataset link.

B.2 Maintenance plan

We shall adhere to a maintenance plan to uphold the integrity of the codebase and ensure the conformity of supplied datasets to requisite standards. In particular, this maintenance plan encompasses:

- Fixing bugs affecting the correctness of our code, whether identified by the community or ourselves;
- Introducing additional variants of federated learning techniques, including alternative methods within the scope of cross-silo federated learning and federated semi-supervised learning methodologies;
- Adding new functional modules, such as privacy protection components.
- Regarding datasets, reviewing potential updates of the datasets referenced in the FedCVD, including but not limited to introducing new tasks or modalities;

C Fed-ECG

C.1 Description

Fed-ECG consists of four datasets: SPH, PTB-XL, SXPH, and G12EC. The order of leads of each dataset is I, II, III, aVR, aVL, aVF, V1, V2, V3, V4, V5, V6. The overview of Fed-ECG is shown in Table 5. Table 6 shows demographics information for four datasets in Fed-ECG.

SPH. The original Shandong Provincial Hospital (SPH) database contains 25,770 12-lead ECG records from 24,666 patients, which were acquired from Shandong Provincial Hospital between 2019/08 and 2020/08. The record length is between 10 and 60 seconds. The sampling frequency is 500 Hz. All ECG records are in full compliance with the AHA standard, which aims for the standardization and interpretation of the electrocardiogram and consists of 44 primary statements and 15 modifiers as per the standard. 46.04% records in this dataset contain ECG abnormalities. Moreover, 14.45% records have multiple diagnostic statements.

Table 5: Overview of the datasets, tasks, metrics and baseline models in FedCVD.

Dataset	Fed-ECG				Fed-ECHO		
Task Type	Multi-label Classification				2D Segmentation		
Input	12-lead ECG Signal				Echocardiogram		
Prediction (y)	Diagnostic Statement				Cardiac Structure Mask		
Data source	SPH	PTB-XL	SXPH	G12EC	CAMUS	ECHONET-DYNAMIC	HMC-QU
Original Patient Size	24,666	18,885	45,152	UNKNOWN	500	10,030	109
Original Sample Size	25,770	21,837	45,152	10,344	1000	20,060	2,349
Preprocessing	Label Alignment				Resizing and Label Alignment		
Patient Size	21,530	16,699	36,272	UNKNOWN	500	10,024	109
Sample Size	22,425	19,019	36,272	6,205	1000	20,048	2,349
Model	ResNet				U-net		
Metrics	Micro F1 / mAP				DICE / Hausdorff distance		
Input Dimension	12×5000				112×112		

PTB-XL. The original PTB-XL database contains 21,837 12-lead ECG records from 18,885 patients of 10 seconds length at the Physikalisch Technische Bundesanstalt (PTB) between October 1989 and June 1996. The original records are resampled to both 100 Hz and 500 Hz. For consistency, we only use the records whose frequency is 500 Hz. Each data is annotated by up to two cardiologists with the SCP-ECG standard.

SXPH. This database contains 12-lead ECGs of 45,152 patients with a 500 Hz sampling rate under the auspices of Chapman University, Shaoxing People's Hospital (Shaoxing Hospital Zhejiang University School of Medicine), and Ningbo First Hospital. The record length is 10 seconds. All records are labeled by professional experts with the SNOMED-CT standard.

G12EC. This Georgia 12-lead ECG Challenge (G12EC) database is provided by the PhysioNet/Computing in Cardiology Challenge 2020. Only 10,344 training data from this database are open to the public. The record length is not longer than 10 seconds with a sample frequency of 500 Hz. All records are labeled with the SNOMED-CT standard as well.

Table 6: Demographics information for Fed-ECG.

Client	Sex	Dataset size	Age	Age Range
Client1	Female	9,502	48.73 ± 15.67	18 - 92
	Male	12,923	50.35 ± 15.49	18 - 95
Client2	Female	8,930	59.80 ± 18.42	3 - 89
	Male	10,089	58.40 ± 15.66	2 - 89
Client3	Female	14,830	58.36 ± 20.11	4 - 89
	Male	21,442	60.28 ± 19.10	4 - 89
Client4	Female	2,668	61.37 ± 16.51	20 - 89
	Male	3,537	61.35 ± 15.04	14 - 89

C.2 License and Ethics

All four databases are open-access. The SPH database is open access at Figshare, while the rest databases are open access at PhysioNet under a Creative Commons Attribution 4.0 International Public License.

The PTB-XL database was supported by the Bundesministerium für Bildung und Forschung (BMBF) through the Berlin Big Data Center under Grant 01IS14013A and the Berlin Center for Machine Learning under Grant 01IS18037I and by the EMPIR project 18HLT07 MedalCare. The EMPIR initiative is cofunded by the European Union's Horizon 2020 research and innovation program and the EMPIR Participating States.

The institutional review board of Shaoxing People's Hospital and Ningbo First Hospital of Zhejiang University approved the study of the SXPH database, granted the waiver application to obtain informed consent, and allowed the data to be shared publicly after de-identification. The requirement for patient consent was waived.

Table 7: Label relationship between original label and ours.

ours	Original Label			
	SPH	PTB-XL	SXPH	G12EC
NORM (Normal)	Normal	Normal	-	-
STACH (Sinus tachycardia)	Sinus tachycardia	Sinus tachycardia	Sinus tachycardia	427084000
SBRAD (Sinus bradycardia)	Sinus bradycardia	Sinus bradycardia	Sinus bradycardia	426177001
SARRH (Sinus arrhythmia)	Sinus arrhythmia	Sinus arrhythmia	-	427393009
PAC (Atrial premature complex(es))	Atrial premature complex(es)	Atrial premature complex	-	-
AFIB (Atrial fibrillation)	Atrial fibrillation	Atrial fibrillation	Atrial fibrillation	164889003
AFLT (Atrial flutter)	Atrial flutter	Atrial flutter	Atrial flutter	164890007
SVTAC (Supraventricular tachycardia)	-	Supraventricular tachycardia	Supraventricular tachycardia	426761007
PVC (Ventricular premature complex)	Ventricular premature complex(es)	Ventricular premature complex	-	164884008
1AVB (First degree AV block)	-	First degree AV block	1 degree atrioventricular block	270492004
	Second-degree AV block, Mobitz type I (Wenckebach)		2 degree atrioventricular block(Type one)	54016002
	Second-degree AV block, Mobitz type II		2 degree atrioventricular block(Type two)	28189009
2AVB (Second degree AV block)	2:1 AV block	Second degree AV block		164903001
	AV block, varying conduction		2 degree atrioventricular block	195042002
	AV block, advanced (high-grade)			284941000119107
3AVB (Third degree AV block)	AV block, complete (third-degree)	Third degree AV block	3 degree atrioventricular block	27885002
	Left anterior fascicular block	Left anterior fascicular block		445118002
LBBB (Left bundle branch block)	Left posterior fascicular block	Left posterior fascicular block	Left bundle branch block	445211001
	Left bundle-branch block	Complete left bundle branch block		164909002
	Incomplete right bundle-branch block	Incomplete right bundle branch block		713426002
RBBB (Right bundle branch block)	Right bundle-branch block	Complete right bundle branch block	Right bundle branch block	59118001
				164907000
LAO/LAE (Left atrial overload/enlargement)	Left atrial enlargement	Left atrial overload/enlargement	-	67741000119109
LVH (Left ventricular hypertrophy)	Left ventricular hypertrophy	Left ventricular hypertrophy	-	164873001
RVH (Right ventricular hypertrophy)	Right ventricular hypertrophy	Right ventricular hypertrophy	-	-
AMI (Anterior myocardial infarction)	Anterior MI	Anterior myocardial infarction	-	-
IMI (Inferior myocardial infarction)	Inferior MI	Inferior myocardial infarction	-	-
ASMI (Anteroseptal myocardial infarction)	Anteroseptal MI	Anteroseptal myocardial infarction	-	-

C.3 DOWNLOAD AND PREPROCESSING

C.3.1 DOWNLOAD

The four datasets can be downloaded using the URLs below:

1. **SPH:** `https://springernature.figshare.com/collections/A_large-scale_multi-label_12-lead_electrocardiogram_database_with_standardized_diagnostic_statements/5779802/1`
2. **PTB-XL:** `https://physionet.org/content/ptb-xl/1.0.3/`
3. **SXPH:** `https://physionet.org/content/ecg-arrhythmia/1.0.0/`
4. **G12EC:** `https://physionet.org/content/challenge-2020/1.0.2/`

C.3.2 PREPROCESSING

Raw 12-lead ECG signals have varying sequence lengths and raw 12-lead ECG signals have varying sequence lengths and annotated standards which must be standardized before FL training. Therefore, we first set a signal length to 10 seconds. We pad the signal with edge value at the edge for those whose length is shorter than 10 seconds and cut off the signal at 10 seconds for those whose length is longer than 10 seconds. Next, we only save the records whose label occurs in at least two databases. Finally, we align the labels of records in different databases. The relationship between the original label and our label is shown in Table7.

C.4 BASELINE, LOSS FUNCTION AND EVALUATION

Baseline Model. We implement a ResNet1d model with 34 layers. The final layer output is passed through a sigmoid function to encode the probability that each label corresponds to one 12-lead ECG signal.

Loss function. The model was directly trained for the Binary CrossEntropy Loss (BCELoss), defined as:

$$\mathrm{BCE}(\mathbf{y}, \hat{\mathbf{y}}) = -[\sum_{i=1}^{n} y_i \log(\hat{y}_i) + \sum_{i=1}^{n}(1 - y_i)\log(1 - \hat{y}_i)] \quad (1)$$

Evaluation Metrics. In multi-label classification for Fed-ECG, the micro F1 score is used as the main metric to evaluate the performance of the model. Given N labels, the micro-precision (P_{micro}) and micro-recall (R_{micro}) are calculated as $P_{\text{micro}} = \frac{\sum_{i=1}^{N} \mathrm{TP}_i}{\sum_{i=1}^{N}(\mathrm{TP}_i + \mathrm{FP}_i)}$ and $R_{\text{micro}} = \frac{\sum_{i=1}^{N} \mathrm{TP}_i}{\sum_{i=1}^{N}(\mathrm{TP}_i + \mathrm{FN}_i)}$, where TP_i is the number of true positives for label i,FP_i is the number of false positives for label i,FN_i is the number of false negatives for label i. The micro F1 score ($F1_{\text{micro}}$) is then calculated as:

$$F1_{\text{micro}} = \frac{2 \cdot P_{\text{micro}} \cdot R_{\text{micro}}}{P_{\text{micro}} + R_{\text{micro}}} \tag{2}$$

For Fed-ECG's Multi-Label Classification task, the Mean Average Precision (mAP) is adopted to measure the classification performance across all labels (including long-tailed labels), calculated by averaging the average precision (AP) for each label, defined as:

$$\text{mAP} = \frac{1}{L} \sum_{i=1}^{L} \sum_{k=1}^{n} P_i(k) \Delta r_i(k) \tag{3}$$

where L is the total number of labels, and AP_i is the average precision for the i-th label, $P_i(k)$ is the precision for label i at the k-th threshold, and $\Delta r_i(k)$ is the change in its recall at the k-th threshold.

C.5 Training Detail

Optimization parameters. We optimize the ResNet1d using SGD optimizer, with a batch size of 32. We train our model for 50 epochs on one NVIDIA A100-PCIE-40GB.

Hyperparameter Search For centralized and local model training, we first conduct a search for optimal learning rates from the set {1e-5, 1e-4, 1e-3, 1e-2, 1e-1} during centralized model training. The learning rate that yields the best micro-F1 score is then used for local model training. For the federated learning strategies, we employ the following hyperparameter grid:

- For clients' learning rates (all strategies): {1e-5, 1e-4, 1e-3, 1e-2, 1e-1}.
- For server size learning rate (Scaffold strategy only): {1e-2, 1e-1, 1.0}.
- For FedProx and Ditto strategies, the parameter μ is selected from {1e-2, 1e-1, 1.0}.
- For FedInit, the parameter β is chosen from {1e-1, 1e-2, 1e-3}.
- For FedSM, the parameters γ and λ are set to values from {0, 0.1, 0.7, 0.9} and {0.1, 0.3, 0.5, 0.7, 0.9}, respectively.
- For FedALA, the parameters layer index, η, threshold, and num_per_loss are fixed at 1, 1.0, 0.1, and 10, respectively, while rand_percent is selected from {5, 50, 80}.

Table 8: Hyperparameters used for the Fed-ECG.

Fed-ECG								
Methods	learning rate	optimizer	learning rate server	mu	beta	lambda	gamma	rand_percent
Central.	0.1	torch.optim.SGD	-	-	-	-	-	-
FedAvg	0.1	torch.optim.SGD	-	-	-	-	-	-
FedProx	0.1	torch.optim.SGD	-	0.01	-	-	-	-
Scaffold	0.1	torch.optim.SGD	1.0	-	-	-	-	-
FedInit	0.1	torch.optim.SGD	1.0	-	0.01	-	-	-
Ditto	0.1	torch.optim.SGD	-	0.01	-	-	-	-
FedSM	0.1	torch.optim.SGD	1.0	-	-	0.1	0	-
FedALA	0.1	torch.optim.SGD	1.0	-	-	-	-	80

Non-IID partition. For the non-IID partition, we first pool the training data from the four clients. Then, we cluster the samples into 10 categories based on the cosine similarity and order them according to the number of samples contained in each category. Next, the sorted samples are divided into 32 shards. finally, 8 random shards are distributed to one client. The label distribution of each client with the non-IID partition is shown in Figure 5.

C.6 Supplementary Experiment Results

We provide additional evaluation metrics here. Table 9 presents an extensive array of evaluation metrics for various federated learning approaches applied to Fed-ECG. The Micro F1-Score (Mi-F1)

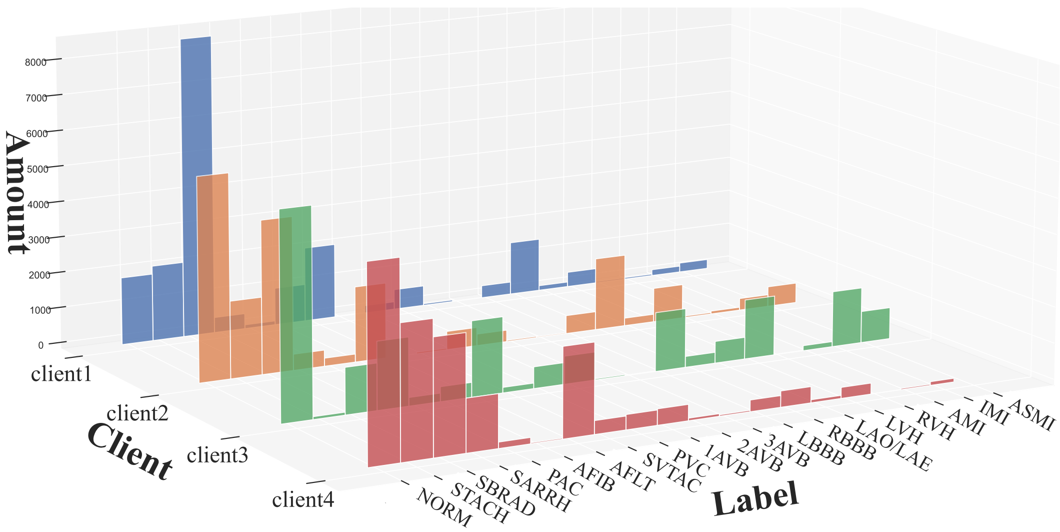


Figure 5: Label non-IID of the Fed-ECG dataset with the artificially non-IID partition, shown as the variation in the number of each label (right axis) across different clients (left axis).

Table 9: The performance of different FL methods on Fed-ECG, with Mi-F1, mAP, and HL representing Micro F1-Score, mean Average Precision score, and Hamming Loss, respectively. All metrics are present in percentage (%). The best results for each configuration are highlighted in **bold**, while the second-best results are underlined.

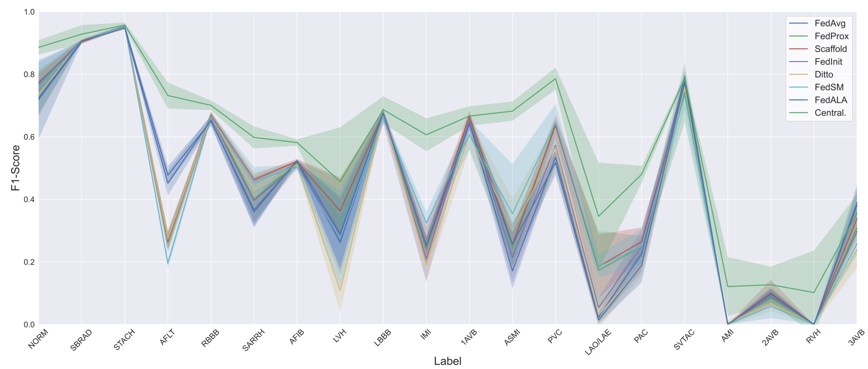
Methods	LOCAL												GLOBAL		
	Client1			Client2			Client3			Client4					
	Mi-F1↑	mAP↑	HL↓	Mi-F1↑	mAP↑	HL↓	Mi-F1↑	mAP↑	HL↓	Mi-F1↑	mAP↑	HL↓	Mi-F1↑	mAP↑	HL↓
Client1	85.8 ±1.9	58.1 ±2.6	1.5 ± 0.2	52.7 ±3.4	37.8 ±2.2	5.8 ± 0.4	61.5 ±1.2	19.8 ±1.2	4.4 ± 0.1	49.8 ±4.2	26.7 ±3.0	6.4 ± 0.6	64.3 ±2.1	32.3 ±2.0	4.1 ± 0.2
Client2	69.9 ±50.0	38.9 ±30.0	3.2 ± 0.1	76.8 ±90.0	55.7 ±50.0	3.1 ± 0.1	26.3 ±80.0	22.7 ±30.0	9.0 ± 0.2	42.2 ±80.0	31.6 ±60.0	8.1 ± 0.1	50.4 ±30.0	35.9 ±70.0	6.1 ± 0.1
Client3	22.7 ±0.2	29.8 ±0.7	8.2 ± 0.0	17.0 ±0.4	27.2 ±0.3	10.3 ± 0.1	88.1 ±0.2	37.7 ±0.4	1.3 ± 0.0	56.9 ±0.4	29.4 ±0.6	5.4 ± 0.1	51.5 ±0.2	32.7 ±0.2	5.5 ± 0.0
Client4	23.7 ±2.0	31.7 ±2.7	8.4 ± 0.9	24.7 ±3.3	30.5 ±1.5	10.1 ± 1.2	61.6 ±5.5	25.3 ±2.1	5.0 ± 1.2	72.3 ±10.2	38.5 ±2.8	4.1 ± 1.8	44.7 ±4.3	29.3 ±2.5	7.0 ± 1.1
FedAvg	69.0 ±10.1	58.5 ±1.2	3.4 ± 1.1	50.3 ±5.3	54.4 ±0.5	6.2 ± 0.7	77.6 ±0.7	37.2 ±0.3	2.5 ± 0.1	66.3 ±0.9	39.5 ±0.5	4.2 ± 0.1	67.9 ±3.8	50.8 ±0.4	3.7 ± 0.5
FedProx	74.0 ±7.5	60.3 ±2.9	2.9 ± 1.0	55.6 ±2.7	56.4 ±0.6	5.5 ± 0.5	73.2 ±1.0	36.0 ±0.8	3.0 ± 0.1	70.2 ±2.3	**43.8 ±1.8**	3.8 ± 0.3	68.8 ±2.6	**52.3 ±0.9**	3.6 ± 0.4
Scaffold	77.5 ±2.6	58.0 ±1.2	2.3 ± 0.2	56.9 ±1.7	55.9 ±0.7	5.2 ± 0.2	73.3 ±1.0	36.2 ±0.6	3.0 ± 0.1	70.7 ±2.9	42.7 ±1.1	3.7 ± 0.3	**70.1 ±0.8**	52.1 ±0.7	**3.4 ± 0.1**
FedInit	73.0 ± 6.6	58.2 ± 0.7	3.1 ± 1.0	54.1 ± 5.2	55.6 ± 1.3	5.9 ± 0.9	73.5 ± 0.5	36.6 ± 0.1	3.0 ± 0.1	67.8 ± 2.0	41.5 ± 1.0	4.1 ± 0.3	68.1 ± 3.0	51.5 ± 0.9	3.8 ± 0.5
Ditto	82.8 ±4.4	**63.1 ±4.2**	1.8 ± 0.4	**74.8 ±1.4**	**58.3 ±0.6**	**3.5 ± 0.2**	86.5 ±1.5	**38.1 ±0.6**	1.5 ± 0.2	**73.4 ±6.7**	42.2 ±4.0	**3.6 ± 0.9**	68.1 ±2.9	48.7 ±1.4	3.6 ± 0.3
FedSM	77.2 ± 7.2	58.8 ± 1.3	2.3 ± 0.6	59.1 ± 4.5	56.4 ± 1.4	5.1 ± 0.5	69.8 ± 0.8	35.0 ± 0.5	3.5 ± 0.1	67.7 ± 3.6	42.9 ± 2.4	4.1 ± 0.4	68.9 ± 2.5	51.2 ± 0.7	3.6 ± 0.3
FedALA	**84.4 ± 4.0**	62.0 ± 7.0	**1.6 ± 0.4**	71.7 ± 5.7	57.1 ± 2.2	3.8 ± 0.6	**88.2 ± 0.1**	37.4 ± 0.2	**1.3 ± 0.0**	66.7 ± 5.9	41.2 ± 2.3	4.4 ± 0.7	67.8 ± 1.9	50.8 ± 1.3	3.7 ± 0.3
Central.	84.9 ±0.5	54.8 ±0.5	1.6 ± 0.1	71.4 ±5.0	55.2 ±2.9	3.8 ± 0.6	84.1 ±1.6	36.5 ±1.1	1.7 ± 0.2	72.2 ±3.7	41.5 ±1.3	3.6 ± 0.3	80.0 ±2.1	63.2 ±2.8	2.3 ± 0.2

and Hamming Loss (HL) serve as indicators of the overall performance, given their insensitivity to long-tail distributions. In contrast, the mean Average Precision score (mAP) provides insight into the average performance across individual labels. In addition, Figure 6 presents the evaluation metrics for each label, encompassing F1 score, precision, and recall, which more clearly demonstrates the impact of the long-tail distribution on each label.

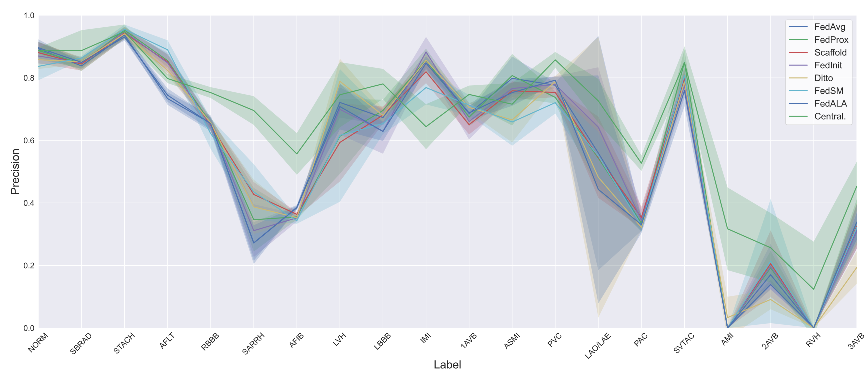
D FED-ECHO

D.1 DESCRIPTION

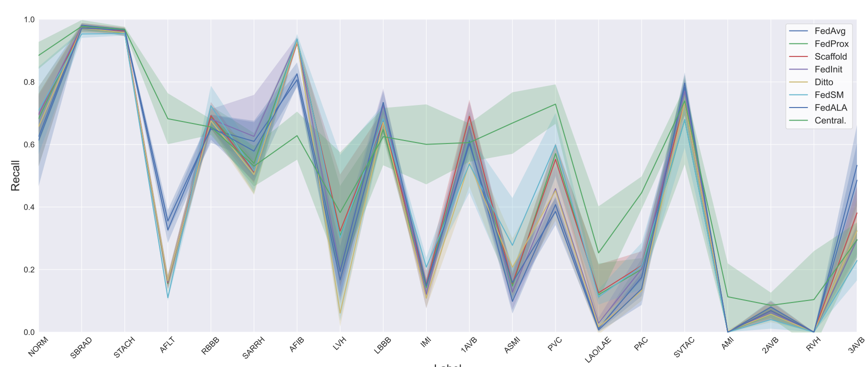
Fed-ECHO consists of three datasets: CAMUS, ECHONET-DYNAMIC, and HMC-QU. The overview of Fed-ECHO is shown in Table 5.



(a) F1-Score for each label



(b) Precision-Score for each label



(c) Recall-Score for each label

Figure 6: Evaluation metrics for each label on Fed-ECG among different FL methods.

CAMUS. This database consists of clinical exams from 500 patients, acquired at the University Hospital of St Etienne (France). All images are labeled with three areas: endocardium of the left ventricle (LV_{Endo}), epicardium of the left ventricle (LV_{Epi}), and left atrium wall (LA). The image size varies from 584×354 to 1945×1181.

ECHONET-DYNAMIC. This database contains 10,0230 echocardiogram videos where two frames are annotated with only LV_{Endo} area. All frames are resized to 112×112.

HMC-QU. This database contains 109 echocardiogram videos collected at the Hamad Medical Corporation Hospital in Qatar. The frames of one cardiac cycle in each video are annotated with LV_{Epi} area. The video frame size varies from 422×636 to 768×1024 while all labels are resized to 224×224.

D.2 License and Ethics

Both CAMUS and HMC-QU datasets are open-access. HMC-QU database requires the user to have a Kaggle account, while the ECHONET-DYNAMIC database requires the user to have a Stanford AIMI account and to accept its agreement. It is licensed under the Stanford University Dataset Research Use Agreement.

D.3 Download and preprocessing

D.3.1 Download

The three datasets can be downloaded using the URLs below:

1. **CAMUS:** `https://humanheart-project.creatis.insa-lyon.fr/database/#collection/6373703d73e9f0047faa1bc8`
2. **ECHONET-DYNAMIC:** `https://echonet.github.io/dynamic/index.html#access`
3. **HMC-QU:** `https://www.kaggle.com/datasets/aysendegerli/hmcqu-dataset/data`

D.3.2 Preprocessing

Raw echocardiograms have varying frame sizes, modalities, and mask labels, which must be standardized before training. Therefore, as a first step, we extract frames that are annotated and store them as images. We then resize them to a common (112×112) shape. Finally, we align the labels of records in different databases. We use 1, 2, 3 representing LV_{Endo}, LV_{Epi} and LA respectively. The samples of Fed-ECHO are shown in Figure7.

(a) Sample from Institution 1. (b) Sample from Institution 2. (c) ample from Institution 3.

Figure 7: Echocardiogram of each institution in Fed-ECHO. LV_{Endo}, LV_{Epi} and LA are shown in red, green and blue respectively.

D.4 BASELINE, LOSS FUNCTION AND EVALUATION

Baseline Model. A U-net architecture is employed in this study, utilizing echocardiographic images as input to forecast masks delineating four distinct cardiac regions. The U-net model represents a conventional convolutional neural network design frequently deployed in the realm of biomedical image segmentation endeavors. Its application is tailored towards semantic segmentation, a process wherein individual pixels within an image are categorized based on semantic content.

Loss function. We use a CrossEntropy Loss (CELoss) for training. Note that, for centralized supervised learning and client training in FedAvg, FedProx, Scaffold, and Ditto strategies, we ignore label with value 0 when calculating CELoss for data from client 2 or 3, since region with label 0 may not be true ground truth in these clients.

Evaluation Metrics. We use the Dice similarity index and 2D Hausdorff distance (d_H) to measure the accuracy of the segmentation output. Dice index is calculated as:

$$\text{DICE}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{2\sum_{i=1}^{n} y_i \hat{y}_i}{\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} \hat{y}_i} \tag{4}$$

The Hausdorff distance is calculated as:

$$\mathrm{d_H}(\mathbf{y}, \hat{\mathbf{y}}) = \max\{d(\mathbf{y}, \hat{\mathbf{y}}), d(\hat{\mathbf{y}}, \mathbf{y})\}, \tag{5}$$

where $d(\mathbf{y}, \hat{\mathbf{y}})$ represents the minimum distance among points at the edge of $\mathbf{y}$ and points at the edge of $\hat{\mathbf{y}}$.

Note that, to better measure the model segmentation performance, for clients 2, and 3, we select only 200 labeled frames for testing.

D.5 TRAINING DETAIL

Optimization parameters. We optimize our model using the SGD optimizer, with a batch size of 32. We train our model for 50 epochs on one NVIDIA A100-PCIE-40GB.

Hyperparameter Search For centralized and local model training, we first explore learning rates from the set {1e-4, 1e-3, 1e-2, 1e-1.5, 1e-1} during centralized model training. The learning rate that achieves the best Dice index is then utilized for local model training. For the federated learning strategies, we employ the following hyperparameter grid:

- For clients' learning rates (all strategies except Fed-Consist): {1e-4, 1e-3, 1e-2, 1e-1.5, 1e-1}.
- For server size learning rate (Scaffold strategy only): {1e-2, 1e-1, 1.0}.
- For FedProx and Ditto strategies, the parameter μ is selected from {1e-2, 1e-1, 1.0}.
- For FedInit, the parameter β is chosen from {1e-1, 1e-2, 1e-3}.
- For FedSM, the parameters γ and λ are set to {0, 0.1, 0.7, 0.9} and {0.1, 0.3, 0.5, 0.7, 0.9}, respectively.
- For FedALA, the parameters layer index, η, threshold, and num_per_loss are fixed at 1, 1.0, 0.1, and 10, respectively, while rand_percent is chosen from {5, 50, 80}.

For Fed-Consist, we introduce Gaussian noise with a variance of 0.1 as augmentation. The learning rates for labeled clients are searched from {1e-2, 1e-3, 1e-4}, while those for unlabeled clients are explored within {1e-3, 1e-4, 1e-5, 5e-6, 1e-6}. The parameter τ is varied from {0.5, 0.7, 0.9}.

Additionally, for FedPSL, we further search the parameters α and β from {1e-0.5, 1e-1, 1e-1.5, 1e-2, 1e-3} and {1e-1, 1e-1.5, 1e-2, 1e-3, 1e-4, 1e-5}, respectively. The optimal values found are $\alpha = 1e-1.5$ and $\beta = 1e-5$.

Table 10: Hyperparameters used for the Fed-ECHO.

Fed-ECHO									
Methods	learning rate	optimizer	learning rate server	mu	beta	lambda	gamma	rand_percent	τ
Central.(sup)	0.1	torch.optim.SGD	-	-	-	-	-	-	-
Central.(ssup)	0.1	torch.optim.SGD	-	-	-	-	-	-	-
FedAvg	0.1	torch.optim.SGD	-	-	-	-	-	-	-
FedProx	0.1	torch.optim.SGD	-	0.1	-	-	-	-	-
Scaffold	0.1	torch.optim.SGD	1.0	-	-	-	-	-	-
FedInit	0.1	torch.optim.SGD	1.0	-	1e-2	-	-	-	-
Ditto	0.1	torch.optim.SGD	-	0.1	-	-	-	-	-
FedSM	0.1	torch.optim.SGD	1.0	-	-	0.1	0	-	-
FedALA	0.1	torch.optim.SGD	1.0	-	-	-	-	5	-
FedPSL	0.1	torch.optim.SGD	1.0	-	1e-5	-	-	-	-
Fed-Consist	0.0001(labeled client) 1e-6(unlabeled client)	torch.optim.SGD	-	-	-	-	-	-	0.9