Beyond full fine-tuning: enhancing generalizability in ECG foundation models' downstream adaptation

Giuliana Monachino[1,2*], Beatrice Zanchi[1,3], Georgiy Farina[1], Francesca Dalia Faraci[1]

[1]Institute of Digital Technologies for Personalized Healthcare - MeDiTech, Department of Innovative Technologies, University of Applied Sciences and Arts of Southern Switzerland, Via la Santa 1, Lugano, 6900, Switzerland.
[2]Institute of Informatics, University of Bern, Neubrückstrasse 10, Bern, 3012, Switzerland.
[3]Department of Quantitative Biomedicine, University of Zurich, Schmelzbergstrasse 26, Zurich, 8091, Switzerland.

*Corresponding author(s). E-mail(s): giuliana.monachino@supsi.ch;

SUPPLEMENTARY INFORMATION

Datasets

This section provides additional details about the datasets used in the pre-training, fine-tuning and out-of-domain evaluation experiments.

Pretraining

Table 1: Number of subjects and 5s-segments in whole dataset, train and validation sets in each of the pre-training datasets.

Dataset	Whole dataset		Train		Val	
	sbj	5s-win	sbj	5s-win	sbj	5s-win
Code-15%	232732	687124	186131	549700	46601	137424
Ningbo	34904	69808	27923	55846	6981	13962
INCART	32	26640	25	21240	7	5400
Off-test	7000	21000	5600	16800	1400	4200
Chapman-Shaoxing	10247	20494	8198	16396	2049	4098
CPSC	6877	20162	5493	16138	1384	4024
CPSC-extra	3453	10196	2741	8175	712	2021
Total	295245	855424	236111	684295	59134	171129

Code-15%

CODE-15% dataset represents a stratified subset (15%) of the larger CODE dataset. It includes 345,779 annotated 12-lead ECG recordings obtained from 233,770 patients. Data collection was conducted by the Telehealth Network of Minas Gerais (TNMG), a public telehealth initiative operating across most municipalities in the state of Minas Gerais, Brazil, during the period from 2010 to 2016.

Chapman-Shaoxing and Ningbo

Chapman University, Shaoxing People's Hospital (Chapman-Shaoxing) dataset and *Ningbo First Hospital (Ningbo)* dataset, were both included in the *PhysioNet/Computing in Cardiology Challenge 2021*. Together, they comprise 45,152 ECG recordings, all made available as training set. Each ECG spans 10 seconds and is recorded at a sampling rate of 500 Hz.

INCART

St Petersburg INCART 12-lead Arrhythmia Database (INCART), also featured in the *PhysioNet/Computing in Cardiology Challenge 2021*, contains 74 annotated ECG signals obtained from 32 Holter monitor recordings. Each segment spans 30 minutes and includes signals from 12 standard leads, all sampled at a rate of 257 Hz.

Off-test

Offline Test Set of ECG Multi-label Classification (Off-test) dataset is the evaluation set of the study titled "Practical arrhythmias detection algorithm for wearable 12-lead ECG via self-supervised learning on large-scale dataset." It consists of 7,000 12-lead ECG recordings captured using wearable devices, each 15 seconds in duration and sampled at 500 Hz. The dataset encompasses 60 rhythm categories, all clinically reviewed and annotated by cardiologists, representing a broad spectrum of both normal and pathological cardiac conditions.

CPSC and CPSC-extra

CPSC and CPSC-extra datasets, made available through the *PhysioNet/Computing in Cardiology Challenge 2021*, originate from the China Physiological Signal Challenge 2018 (CPSC2018). These datasets include two collections comprising 6,877and 3,453 12-lead ECG recordings, respectively. The durations of these recordings vary between 6 and 60 seconds, with each sampled at a frequency of 500 Hz.

Downstream adaptation and in-domain evaluation

CinC2017

The 2017 PhysioNet/CinC Challenge (CinC2017) dataset comprises single-lead ECG recordings collected using the AliveCor device and provided by the company for the challenge. The training set includes 8,528 recordings ranging from 9 to over 60 seconds in duration, while the test set contains 3,658 similarly sized recordings. All ECGs were sampled at 300 Hz and preprocessed with bandpass filtering by the device. The dataset contains four different labels: normal sinus rhythm, atrial fibrillation (AF), other rhythms, and noisy recordings.
In our study, only the training set of the CinC2017 dataset has been used, splitting it on a subject basis into training, validation, and test sets according to an 80:10:10 ratio and label stratification. For the binary AF detection task, we considered AF versus not AF (including normal sinus rhythm, other rhythms, and noisy recordings).

AFDB

MIT-BIH Atrial Fibrillation Database (AFDB) includes 23 long-term ECG recordings lasting 10 hours and recorded from subjects with atrial fibrillation, mostly paroxysmal. Each ECG signal is sampled at 250 Hz with 12-bit resolution over a range of ± 10 millivolts, and two leads (lead I and II) are available for each recording. They have been recorded at Boston's Beth Israel Hospital (now the Beth Israel Deaconess Medical Center) with ambulatory ECG recorders and released in November 2000. Rhythm was manually annotated considering four classes: atrial fibrillation, atrial flutter, AV junctional rhythm, and all other rhythms. Beat annotations obtained using an automated detector, and for some recordings also manually corrected, are also available.
In our study, to match the in-domain dataset, only lead I has been selected, and the labels have been assigned for binary classification: AF versus not AF (including atrial flutter, AV junctional rhythm, and all other rhythms).

PTB-XL

PTB-XL dataset contains 21'799 clinical 12-lead ECG recordings, each 10 seconds long, from 18'869 patients aged 0 to 95 years (median age 62), with a balanced gender distribution. The data was collected using Schiller AG devices between 1989 and 1996, and with a sampling frequency of 500 Hz. Each ECG was annotated by up to two cardiologists with a report string, which was converted into a set of standardized SCP-ECG diagnostic, form, and rhythm statements, and validated by technical experts. Along with the raw waveform data, the dataset includes patient metadata such as age, sex, weight, and height. The dataset is valuable for its broad representation of cardiac conditions, including many healthy controls, and comprehensive annotations with unique ECG and patient identifiers.
In our work, we used the 22 diagnostic statements (as reported in [**?**]) for the cardiac diagnosis detection task, their association in 5 superclasses (as reported in [**?**]) for the cardiac diagnostic superclasses detection task, and the atrial fibrillation label for the AF detection task. We split the dataset on a subject basis into training, validation, and test sets according to an 80:10:10 ratio and label stratification (based on the 22 diagnostic statements).

Out-of-domain evaluation

GEORGIA

Georgia dataset, from *PhysioNet/Computing in Cardiology Challenge 2021*, represents a unique demographic of the Southeastern United States. It includes 20,672 ECG recordings, with 10,344 used for training, 5,167 for validation, and 5,161 for testing. Each ECG lasts between 5 and 10 seconds and is sampled at a frequency of 500 Hz.
We exploited just the training set of this dataset since it is the only one made available. It has been split on a subject basis into training, validation, and test sets according to an 80:10:10 ratio and label stratification. We used the 22 diagnostic statements (as reported in [**?**]) for the cardiac diagnosis detection task, and the atrial fibrillation label for the AF detection task.

VFDB

MIT-BIH malignant ventricular ectopy database (VFDB) contains 22 30-minute 2-leads ECG recordings, sampled at 250 Hz, of subjects who experienced episodes of sustained ventricular tachycardia, ventricular flutter, and ventricular fibrillation. Rhythm annotations are provided, with the time reference of each rhythm change. For our binary LTA detection task, we grouped ventricular tachycardia, ventricular flutter, and ventricular fibrillation rhythms in the LTA class and all the other rhythms in the non-LTA class. In this work, we selected only lead II for consistency with the CUDB database.

CUDB

Creighton University ventricular tachyarrhythmia database (CUDB) contains 35 8.5-minute single-lead ECG recordings, sampled at 250 Hz, of subjects who experienced episodes of ventricular fibrillation. Rhythm annotations were provided, with the time reference of each rhythm change. For our binary LTA detection task, we included ventricular fibrillation in the LTA class and all the other rhythms in the non-LTA class.

Experimental settings

Fine-tuning and supervised training experiments have been conducted under the same settings. Adam optimizer with exponential decay has been used with initial learning rate $lr = 1e - 5$ and decay factor $\gamma = 0.97$. Early stopping has been applied based on validation loss, with a maximum of 50 training epochs, a patience of 10, and a minimum improvement threshold of $1e - 3$. For *CinC2017* and *PTB-XL* datasets, batch size has been fixed to 128, 32, and 8 for 100%, 10%, and 1% subsets, respectively. For *VFDB* dataset, instead, given its smaller size, a batch size of 32 has been used.

Given the high class imbalance in all the datasets and tasks, the following data sampling strategies have been applied for generating the training batches during fine-tuning. First, random uniform *class sampling* is applied, to have on average 50% of AF and 50% of not AF samples in each batch. Then, a recording containing the selected label is identified according to the probability $P(R) = \alpha P_1(R) + (1 - \alpha)P_2(R)$, where $P^1(R)$ is the uniform distribution probability over the suitable recordings, while $P^2(R)$ is the probability of sampling a recording proportionally to its length. α is set to 0.5 to assign equal weights to the two probabilities. Finally, a 5s segment is selected from the identified recording with uniform probability. For each epoch, $Nb = Tr/N$ batches are generated and used, with Tr equal to the number of 5s segments in the train set, and N equal to the batch size.

Results

Table 2: Results obtained with the first scenario (AF detection using CinC2017 dataset). The AUROC evaluated on in-domain (CinC2017 test set) and out-of-domain (AFDB and GEORGIA datasets), and the generalization gap (difference between training and test performance on CinC2017) are reported.

	Method	CinC2017 (test)	Gen. gap	AFDB	GEORGIA
AF detection CinC2017 - 100%	Supervised S	89.94 ± 0.43	+9.67	81.99 ± 1.13	78.72 ± 1.34
	Supervised	87.78 ± 1.21	+11.14	82.54 ± 1.79	79.24 ± 2.31
	Full fine-tuning	95.27 ± 0.34	+4.69	95.04 ± 1.37	86.33 ± 2.51
	Partial fine-tuning 1B	96.26 ± 0.16	+2.94	98.83 ± 0.04	90.55 ± 0.31
	LayerNorm tuning	**96.42 ± 0.09**	+1.55	**99.12 ± 0.09**	92.73 ± 0.37
	BitFit	96.39 ± 0.11	+1.73	**99.12 ± 0.13**	92.73 ± 0.39
	Linear Probing	96.37 ± 0.05	**+0.79**	98.99 ± 0.09	**93.21 ± 0.26**
AF detection CinC2017 - 10%	Supervised S	82.12 ± 0.73	+17.69	69.86 ± 0.75	75.53 ± 0.95
	Supervised	80.76 ± 1.17	+18.52	70.38 ± 1.48	71.96 ± 1.65
	Full fine-tuning	94.51 ± 0.94	+5.48	94.96 ± 1.50	89.34 ± 1.05
	Partial fine-tuning 1B	**95.17 ± 0.08**	+4.60	98.63 ± 0.05	90.52 ± 0.27
	LayerNorm tuning	93.24 ± 0.37	+5.27	99.15 ± 0.08	91.24 ± 0.37
	BitFit	93.38 ± 0.32	+5.20	98.98 ± 0.14	91.25 ± 0.39
	Linear Probing	95.06 ± 0.21	**+2.92**	**99.24 ± 0.08**	**92.09 ± 0.28**
AF detection CinC2017 - 1%	Supervised S	70.92 ± 1.15	+28.99	61.29 ± 1.88	53.95 ± 1.06
	Supervised	67.12 ± 1.65	+32.12	61.98 ± 0.76	50.15 ± 2.37
	Full fine-tuning	89.27 ± 0.67	+10.73	93.67 ± 4.53	88.10 ± 2.27
	Partial fine-tuning 1B	91.34 ± 0.33	+8.55	**98.56 ± 0.18**	89.54 ± 0.55
	LayerNorm tuning	91.98 ± 0.93	+6.68	98.53 ± 0.88	89.84 ± 1.55
	BitFit	92.00 ± 0.74	+6.69	98.55 ± 0.97	**90.01 ± 1.74**
	Linear Probing	**92.29 ± 1.45**	**+5.23**	98.47 ± 0.56	89.83 ± 0.97

Table 3: Results obtained with the second scenario (AF detection using PTB-XL dataset). The AUROC evaluated on in-domain (PTB-XL test set) and out-of-domain (AFDB and GEORGIA datasets), and the generalization gap (difference between training and test performance on PTB-XL) are reported.

	Method	CinC2017 (test)	Gen. gap	AFDB	GEORGIA
	Supervised S	96.30 ± 0.34	+3.64	83.84 ± 3.67	88.96 ± 0.35
	Supervised	94.98 ± 0.57	+4.71	82.91 ± 3.33	88.87 ± 0.55
AF detection	Full fine-tuning	97.38 ± 0.22	+2.60	98.98 ± 0.17	91.66 ± 0.30
	Partial fine-tuning 2B	97.87 ± 0.27	+2.08	99.45 ± 0.05	92.09 ± 0.15
PTB-XL - 100%	LayerNorm tuning	98.18 ± 0.13	+1.44	99.52 ± 0.04	92.78 ± 0.24
	BitFit	**98.25 ± 0.11**	**+1.41**	**99.54 ± 0.04**	92.71 ± 0.29
	Linear Probing	97.86 ± 0.10	+1.63	99.44 ± 0.03	**92.83 ± 0.15**
	Supervised S	91.45 ± 0.74	+8.52	62.09 ± 2.93	80.87 ± 1.59
	Supervised	90.51 ± 1.44	+9.42	67.45 ± 1.72	80.82 ± 1.00
AF detection	Full fine-tuning	97.01 ± 0.38	+2.99	99.09 ± 0.13	91.34 ± 0.73
	Partial fine-tuning 2B	96.97 ± 0.14	+2.99	99.39 ± 0.05	91.32 ± 0.14
PTB-XL - 10%	LayerNorm tuning	97.50 ± 0.09	+1.27	**99.52 ± 0.03**	**92.78 ± 0.14**
	BitFit	**97.63 ± 0.20**	+1.33	99.47 ± 0.08	92.73 ± 0.17
	Linear Probing	97.34 ± 0.07	**+0.97**	99.46 ± 0.06	92.74 ± 0.14
	Supervised S	72.97 ± 6.93	+23.87	55.98 ± 1.08	61.88 ± 4.08
	Supervised	62.28 ± 9.10	+35.87	58.86 ± 1.21	59.05 ± 3.38
AF detection	Full fine-tuning	97.40 ± 0.24	+2.60	98.69 ± 0.13	91.22 ± 0.61
	Partial fine-tuning 2B	**97.46 ± 0.16**	**+2.54**	**99.17 ± 0.09**	**91.49 ± 0.11**
PTB-XL - 1%	LayerNorm tuning	95.95 ± 0.69	+2.74	98.83 ± 0.49	90.79 ± 0.65
	BitFit	96.00 ± 0.65	+2.84	98.94 ± 0.38	90.93 ± 0.61
	Linear Probing	95.54 ± 0.91	+2.64	98.40 ± 0.83	90.23 ± 0.89

Table 4: Results obtained with the third scenario (LTA detection using VFDB dataset). The AUROC evaluated on in-domain (VFDB test set) and out-of-domain (CUDB dataset), and the generalization gap (difference between training and test performance on VFDB) are reported.

	Method	VFDB (test)	Gen. gap	CUDB
	Supervised S	98.80 ± 0.19	+1.20	92.90 ± 0.69
	Supervised	98.86 ± 0.34	+1.13	93.93 ± 1.17
LTA detection	Full fine-tuning	99.06 ± 0.42	+0.94	98.12 ± 0.29
	Partial fine-tuning 1B	**99.45 ± 0.04**	**+0.46**	97.36 ± 0.15
VFDB - 100%	LayerNorm tuning	98.59 ± 0.17	+0.79	98.10 ± 0.07
	BitFit	98.71 ± 0.19	+0.71	**98.18 ± 0.12**
	Linear Probing	95.51 ± 0.19	+2.69	95.53 ± 0.12

Table 5: Results obtained with the fourth scenario (Cardiac diagnosis with 5 superclasses using PTB-XL dataset). The macro AUROC evaluated in-domain (PTB-XL test set), and the generalization gap (difference between training and test performance on PTB-XL) are reported.

	Method	PTB-XL (test)	Gen. gap
Cardiac diagnosis (5 superclasses) PTB-XL - 100%	Supervised S	87.24 ± 0.30	+5.64
	Supervised	87.76 ± 0.19	+7.12
	Full fine-tuning	**87.94 ± 0.37**	+11.42
	Partial fine-tuning 3B	87.41 ± 0.15	+7.83
	LayerNorm tuning	84.08 ± 0.11	+2.00
	BitFit	84.79 ± 0.14	+1.88
	Linear Probing	78.97 ± 0.14	**+0.95**
Cardiac diagnosis (5 superclasses) PTB-XL - 10%	Supervised S	80.05 ± 0.91	+15.80
	Supervised	81.47 ± 0.26	+14.70
	Full fine-tuning	**83.83 ± 0.47**	+16.15
	Partial fine-tuning 6B	82.41 ± 0.41	+17.47
	LayerNorm tuning	78.19 ± 0.13	+12.44
	BitFit	78.86 ± 0.12	+12.39
	Linear Probing	75.21 ± 0.25	**+12.12**
Cardiac diagnosis (5 superclasses) PTB-XL - 1%	Supervised S	74.30 ± 0.35	+23.19
	Supervised	73.97 ± 0.47	+23.09
	Full fine-tuning	**76.98 ± 0.54**	+23.01
	Partial fine-tuning 6B	75.77 ± 0.47	+24.21
	LayerNorm tuning	69.11 ± 1.33	+21.48
	BitFit	69.37 ± 1.28	+21.46
	Linear Probing	68.13 ± 1.32	**+21.24**

Table 6: Results obtained with the fifth scenario (Cardiac diagnosis with 22 classes using PTB-XL dataset). The macro AUROC evaluated in-domain (PTB-XL test set) and out-of-domain (GEORGIA dataset), and the generalization gap (difference between training and test performance on PTB-XL) are reported.

	Method	PTB-XL (test)	Gen. gap	GEORGIA
Cardiac diagnosis (22 classes) PTB-XL - 100%	Supervised S	85.91 ± 0.52	+7.46	76.61 ± 0.24
	Supervised	86.52 ± 0.18	+9.74	76.42 ± 0.35
	Full fine-tuning	88.82 ± 0.53	+10.51	77.82 ± 0.57
	Partial fine-tuning 6B	**89.05 ± 0.32**	+9.18	**79.09 ± 0.26**
	LayerNorm tuning	82.60 ± 0.15	+1.37	76.37 ± 0.24
	BitFit	82.78 ± 0.18	**+1.22**	76.69 ± 0.20
	Linear Probing	80.30 ± 0.06	+2.14	74.36 ± 0.10
Cardiac diagnosis (22 classes) PTB-XL - 10%	Supervised S	75.91 ± 0.81	+18.67	69.65 ± 0.55
	Supervised	75.62 ± 0.78	+18.59	69.28 ± 0.33
	Full fine-tuning	84.41 ± 0.73	+15.47	76.11 ± 0.63
	Partial fine-tuning 6B	**84.70 ± 0.30**	+14.88	**77.48 ± 0.48**
	LayerNorm tuning	76.71 ± 0.52	+7.62	70.11 ± 0.33
	BitFit	76.87 ± 0.61	**+7.33**	70.18 ± 0.39
	Linear Probing	75.33 ± 0.46	+7.58	69.02 ± 0.31
Cardiac diagnosis (22 classes) PTB-XL - 1%	Supervised S	64.36 ± 0.70	+21.86	60.91 ± 0.80
	Supervised	63.36 ± 0.79	+29.33	61.03 ± 0.78
	Full fine-tuning	75.04 ± 0.79	+24.85	72.06 ± 0.48
	Partial fine-tuning 5B	**76.24 ± 0.46**	+23.61	**72.32 ± 0.44**
	LayerNorm tuning	65.75 ± 0.89	+19.81	61.38 ± 0.95
	BitFit	65.97 ± 0.83	+19.87	61.51 ± 1.00
	Linear Probing	65.06 ± 0.93	**+19.45**	60.92 ± 0.90

Table 7: Per-subject results obtained on AFDB dataset in the first scenario (AF detection using CinC2017 dataset). Mean and standard deviation of the AUROC across subjects are reported.

	Method	AFDB
CinC2017 AF detection - 100%	Supervised S	84.41 ± 15.88
	Supervised	85.18 ± 16.14
	Full fine-tuning	96.60 ± 3.08
	Partial fine-tuning 1B	97.67 ± 3.21
	LayerNorm	99.04 ± 1.68
	BitFit	98.91 ± 1.86
	Linear probing	99.19 ± 1.30
CinC2017 AF detection - 10%	Supervised S	69.43 ± 25.93
	Supervised	76.88 ± 20.65
	Full fine-tuning	95.87 ± 6.32
	Partial fine-tuning 1B	98.10 ± 2.62
	LayerNorm	98.83 ± 1.58
	BitFit	98.66 ± 1.79
	Linear probing	99.16 ± 1.16
CinC2017 AF detection - 1%	Supervised S	64.99 ± 17.83
	Supervised	70.59 ± 18.20
	Full fine-tuning	93.26 ± 7.58
	Partial fine-tuning 1B	97.73 ± 4.82
	LayerNorm	98.26 ± 1.99
	BitFit	98.28 ± 1.99
	Linear probing	98.29 ± 1.96

Table 8: Per-subject results obtained on AFDB dataset in the second scenario (AF detection using PTB-XL dataset). Mean and standard deviation of the AUROC across subjects are reported.

	Method	AFDB
PTB-XL AF detection - 100%	Supervised S	86.73 ± 16.18
	Supervised	85.67 ± 15.54
	Full fine-tuning	98.65 ± 1.74
	Partial fine-tuning 2B	98.53 ± 2.48
	LayerNorm	99.26 ± 1.04
	BitFit	99.20 ± 1.17
	Linear probing	99.24 ± 1.09
PTB-XL AF detection - 10%	Supervised S	70.89 ± 19.23
	Supervised	74.74 ± 17.98
	Full fine-tuning	98.51 ± 2.19
	Partial fine-tuning 2B	98.27 ± 2.96
	LayerNorm	99.12 ± 1.33
	BitFit	99.00 ± 1.48
	Linear probing	99.14 ± 1.31
PTB-XL AF detection - 1%	Supervised S	59.37 ± 17.08
	Supervised	61.00 ± 16.88
	Full fine-tuning	98.34 ± 1.90
	Partial fine-tuning 2B	98.99 ± 1.52
	LayerNorm	98.47 ± 1.70
	BitFit	98.51 ± 1.62
	Linear probing	98.00 ± 2.29

Table 9: Per-subject results obtained on VFDB and CUDB dataset in the third scenario (AF detection using VFDB dataset). Mean and standard deviation of the AUROC across subjects are reported.

	Method	VFDB (test)	CUDB
VFDB - LTA 100%	Supervised S	99.30 ± 0.53	96.61 ± 4.73
	Supervised	99.21 ± 1.00	96.43 ± 5.55
	Full fine-tuning	98.36 ± 1.46	98.64 ± 2.53
	Partial fine-tuning 1B	99.20 ± 0.99	98.40 ± 3.81
	LayerNorm	98.41 ± 2.36	98.64 ± 3.28
	BitFit	98.68 ± 1.78	98.64 ± 3.25
	Linear probing	92.58 ± 9.85	96.44 ± 8.99