

Exercise Heart Rate Estimation Using Range-Covering Deep Learning on Wrist Photoplethysmography Signals

Anonymous Authors
Paper ID assigned by CMT

Abstract—Deep networks that estimate heart rate (HR) from wrist photoplethysmography (PPG) are mostly trained on PPG-DaLiA, a daily-life dataset in which only 3.4% of the windows exceed 140 beats per minute (bpm). During running, these networks underestimate high heart rates. We propose Range-Covering Deep Learning (RCDL), which adapts a small pre-trained estimator to the missing range. RCDL gives the encoder a head that predicts a distribution over HR bins from the embedding and the signal spectrum. It is trained for a fixed number of epochs on PPG-DaLiA mixed with 59 minutes of real treadmill recordings, and its outputs are tracked over time. We selected RCDL on three development datasets and tested it with pre-registered rules on a held-out dataset and a smartwatch. On the held-out dataset, RCDL reduced the mean absolute error from 18.0 to 11.9 bpm against the same pipeline trained without the treadmill pool. On all five external datasets, the reduction was 19% to 40%, and the error on held-out PPG-DaLiA subjects did not change detectably. Ablations show that the pool and the head each reduce the error. With 76k parameters, RCDL is smaller than BeliefPPG but less accurate on most datasets.

Index Terms—Data augmentation, domain shift, fine-tuning, heart rate estimation, photoplethysmography (PPG), wearable sensors

I. INTRODUCTION

Wrist-worn photoplethysmography (PPG) provides continuous heart rate (HR) monitoring, but motion artifacts make the estimate unreliable during exercise. Deep networks that fuse PPG with acceleration are now the standard approach [1]–[5], and most of them are trained and compared on PPG-DaLiA (DaLiA below) [1], a dataset of daily activities.

DaLiA rarely contains high heart rates. Only 3.4% of its 64 697 windows exceed 140 beats per minute (bpm), but 36%–63% of the evaluated windows of four newer datasets do (Fig. 1). The released PULSE teacher [3] reaches 4.03 bpm on DaLiA and 35.3 bpm during jogging and running on GalaxyPPG [6], partly because its estimates stay below the reference above its training range. In the terms of deep imbalanced regression [7], the high-HR part of the target space is almost empty in training but dense at test time. TimePPG reduced the error of its worst-tracked DaLiA subject by fine-tuning on part of that subject’s recording [8].

Synthetic range augmentation plays training windows faster or slower and scales the label [4], [5], [9] or uses such windows for pretraining [10]. Multi-dataset training adds data from other protocols. BeliefPPG, for example, is trained and evaluated on eight datasets, among them the treadmill

recordings we use below, and it tracks a predicted distribution over HR bins [5]. We ask instead whether an estimator already trained on DaLiA can be adapted to the range it has not seen, with an hour of labeled data and without labels from the test subject.

We propose Range-Covering Deep Learning (RCDL). RCDL fine-tunes the pretrained PULSE student [3] with a head that predicts a distribution over HR bins. It mixes a small pool of real treadmill recordings into the DaLiA batches, trains for a fixed budget without selecting checkpoints on DaLiA, and tracks the predicted distributions. Our contributions are as follows.

- We describe the failure of DaLiA-trained estimators during running as a missing target range and propose RCDL, a pipeline of 76k parameters that adapts a pretrained encoder to this range.
- We select RCDL among several variants on three development datasets and test it with pre-registered rules on a held-out dataset and a smartwatch. On the held-out dataset, it reduced the error by 34%.
- We ablate each component one factor at a time and audit the evaluation metric against a standard accuracy criterion and constant predictors.

II. RELATED WORK

A. Deep Learning for Wrist Heart Rate Estimation

DeepPPG [1] introduced PPG-DaLiA and a convolutional network on spectra. TimePPG [8] searched temporal convolutional networks for microcontrollers. PULSE fuses PPG and acceleration with cross-attention [2], and a smaller student network is distilled from it [3]. KID-PPG adds an adaptive motion filter [4]. BeliefPPG trains a network against Gaussian targets over HR bins and tracks its output with belief propagation under a prior on the logarithmic change of HR [5]. Pulse-PPG is a foundation model pretrained on field PPG [11].

B. Imbalanced Targets and Augmentation

Deep imbalanced regression studies continuous targets whose distribution is skewed or has empty regions [7]. SMOTER oversamples rare target values [12], and C-Mixup mixes samples with similar labels to improve regression under distribution shift [13]. Label shift, a change of the label distribution between datasets, is usually corrected by re-weighting [14]. For PPG, AugmentPPG speeds up or slows down DaLiA

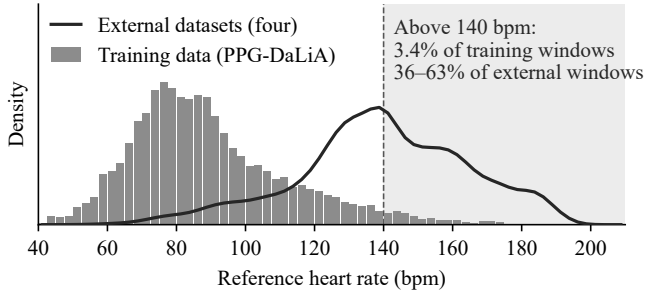


Fig. 1. Reference HR of the training windows (PPG-DaLiA) and of the evaluated windows of four external datasets. The shaded range is almost absent from training.

windows and scales the label [9]. KID-PPG uses a twofold playback [4], EnhancePPG applies such augmentation during self-supervised pretraining [10], and BeliefPPG trains with random time stretches [5]. RCDL differs from this work in two ways. It adapts an already trained estimator with a small pool aimed at the uncovered range, instead of augmenting all training windows or training on many datasets. It also does not select checkpoints on the training distribution. Unlike TimePPG, it needs no labels from the test subject.

III. RANGE-COVERING DEEP LEARNING

Fig. 2 shows the four steps of RCDL.

A. Problem Setting

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n_D}$ be the training set, where x_i is an 8-s window of PPG and three-axis acceleration and y_i is the reference HR. We call a range of HR values *uncovered* when the training density $p_{\mathcal{D}}(y)$ is close to zero there. In this work, the uncovered range lies above 140 bpm. It holds 3.4% of the DaLiA windows but 36%–64% of the evaluated windows of the external datasets (Table I). As a working model, we treat the change between datasets as a label shift [14], in which $p(x | y)$ stays fixed and only $p(y)$ changes. An estimator trained on $\mathcal{D}$ then approximates the posterior

$$p_{\mathcal{D}}(y | x) \propto p(x | y) p_{\mathcal{D}}(y). \quad (1)$$

Where $p_{\mathcal{D}}(y)$ is close to zero, the posterior has little mass even when the signal x indicates a high rate, and the estimate stays near the upper edge of the covered range. Re-weighting cannot correct this, because it needs training samples wherever the test labels occur [14]. RCDL therefore adds samples to the uncovered range. The model is approximate, because running also changes the motion content of x (Section VI).

B. Range-Covering Pool and Mixed Batches

The pool $\mathcal{P} = \{(x_j, y_j)\}_{j=1}^{n_P}$ consists of the 12 treadmill recordings of the IEEE Signal Processing Cup 2015 (SPC 2015) [15]. They last 59 minutes and give $n_P = 1726$ windows, 52% of them above 140 bpm. Their PPG is the mean of the two standardized PPG channels, and their acceleration is standardized with the statistics of the DaLiA fold. The labels are the official ECG-derived HR on the same 8-s windows

with a 2-s hop. Because the pool is small, each pool window is repeated $r = 10$ times in the set from which batches are drawn. The expected share of pool windows in a batch is then

$$\pi = \frac{r n_P}{n_D + r n_P} \approx 26\% \quad (2)$$

for the $n_D = 46\,624$ to $51\,690$ windows of the 11 or 12 training subjects of a fold (π between 25% and 27%).

C. Encoder and Distribution Head

The encoder is the PULSE student [3]. It has 27 241 parameters and was released as 15 checkpoints, one per leave-one-subject-out fold of DaLiA. It maps a window, resampled to 32 Hz, to a 32-dimensional embedding. The distribution head replaces the scalar output. It takes the embedding and a fixed log-frequency spectrum of the four input channels, in 158 bins with 32 bins per octave from 15 to 460 bpm. One hidden layer of 64 units maps them to logits over 96 log-spaced HR bins from 30 to 240 bpm. The head has 48 800 parameters, so RCDL has 76 041, against 131 825 for the PULSE teacher and 138k for BeliefPPG [5]. As in BeliefPPG, the head is trained with the cross-entropy against a Gaussian target centered on the reference HR, here with a standard deviation of 1.5 bins. The estimate of a window is the soft-argmax over the most likely bin and the three bins on each side.

D. Fixed-Budget Training

The encoder of each fold starts from its released checkpoint and is fine-tuned with Adam at 10^{-4} . The head is trained from scratch at 10^{-3} . Batches of 256 windows are drawn uniformly from the DaLiA training subjects of the fold and the repeated pool. Training runs for 20 epochs of $n_D/256$ steps, and the final weights are kept. We do not select checkpoints on DaLiA validation subjects, because DaLiA has almost no labels in the uncovered range and cannot reward an improvement there.

E. Tracking

Consecutive windows overlap by 6 s, and the HR changes little from one window to the next. The tracker combines the per-window distributions with a transition prior that, as in BeliefPPG [5], depends only on the change on the logarithmic HR axis. Our prior is a Laplace distribution over the bin distance with a width of 0.5 bins per 2-s step, truncated at 14 bins. A fixed-lag forward-backward pass with a lag of two windows (4 s) gives the posterior of each window, and its mean on the log-frequency axis is the estimate. Because the prior does not depend on the absolute HR, the tracker is shift-invariant and adds temporal continuity without favoring the training range. Untracked outputs use the post-processing of PULSE, which clips each estimate to within 10% of the mean of the previous ten.

IV. EXPERIMENTAL SETUP

A. Datasets and Preprocessing

Table I lists the datasets. ScientISST MOVE [16], GalaxyPPG [6], and PPG-Sport [17] are the development datasets. WEEE [18] was held out until the pipeline and

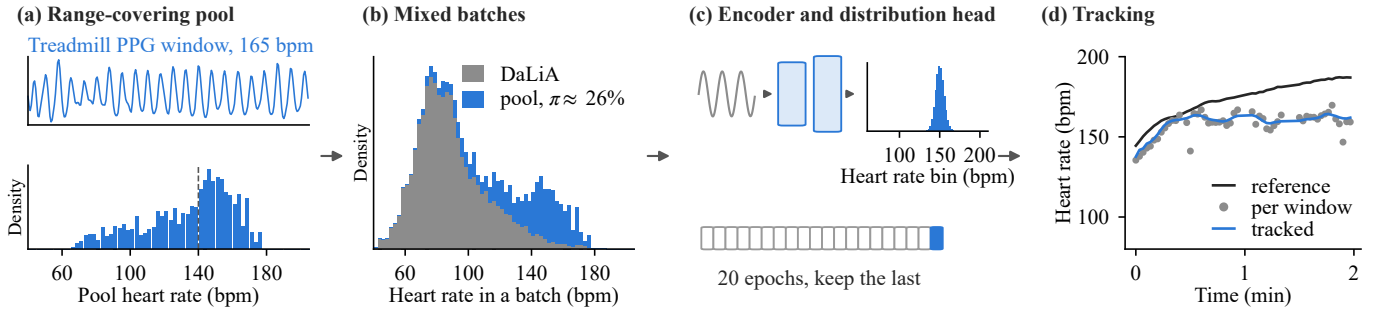


Fig. 2. Overview of RCDL: (a) range-covering pool of real treadmill windows, (b) batches that mix DaLiA and pool windows at a fixed share, (c) pretrained encoder with a distribution head over HR bins, trained for a fixed number of epochs, and (d) reference, per-window, and tracked HR of one GalaxyPPG participant during running.

TABLE I
DATASETS

Dataset	Device	Subjects	>140 bpm	Role
PPG-DaLiA [1]	E4	15	3.4%	training
SPC 2015 [15]	wrist PPG	12 rec.	52%	pool
ScientISST MOVE [16]	E4	16	59%	development
GalaxyPPG [6]	E4	18	63%	development
PPG-Sport [17]	EmbracePlus	30	44%	development
WEEE [18]	E4	14	36%	held-out
Galaxy Watch [6]	Galaxy Watch 5	17	64%	device test

Subjects: evaluated subjects. Galaxy Watch: the watch channel of GalaxyPPG, worn by the same participants.

its analysis scripts were fixed. The Galaxy Watch channel of GalaxyPPG was recorded on the other wrist of the same participants in the same sessions, and it tests a change of device within the same cohort. ScientISST, GalaxyPPG, and WEEE use the Empatica E4 of DaLiA. For the E4 datasets, the reference HR is the consensus of six electrocardiogram (ECG) R-peak detectors [19], accepted when at least four agree within 2 bpm. PPG-Sport provides ECG-derived HR. All windows are 8 s long with a hop of 2 s. We evaluate the jogging, running, and treadmill windows, and all valid windows for ScientISST. PPG and acceleration are standardized with the statistics of the DaLiA fold. The EmbracePlus and Galaxy Watch PPG are standardized per subject and inverted in polarity, and the Galaxy Watch PPG is band-pass filtered at 0.5–4 Hz, as in the dataset paper.

B. Pipeline Selection and Protocol

We fixed the pool and its share before the first experiment, and the fixed budget replaced early stopping after our first run, in which early stopping on DaLiA returned the pretrained model in most folds. The head, the tracker, and the encoder were chosen on the development datasets (Section V-B). The best variant, RCDL, was then tested on WEEE and the Galaxy Watch. Its first pre-registered test on WEEE required it to beat both the no-pool control and the scalar student followed by a 4-s smoother. The pool gain was significant (5.7 bpm, Holm $p = 0.040$), but the gain over the smoothed scalar student was not (5.3 bpm, Holm $p = 0.078$). A second pre-registered rule, written after that test, adopted RCDL if the

pool improved it significantly on WEEE (Holm over WEEE and the watch). RCDL also could not be significantly worse than the scalar teacher with the pool there. Both held (Holm $p = 0.017$ and 0.004). Every hypothesis, null, and decision rule was committed to a version-controlled repository before its run. Test subject k is evaluated with the model of fold $(k \bmod 15) + 1$. Each configuration is trained with two seeds, and their estimates are averaged per window. The control, called the no-pool control below, is the same pipeline trained without the pool. Arms are compared with paired two-sided Wilcoxon signed-rank tests over subjects, with a Holm correction for each family of tests and subject-level bootstrap 95% confidence intervals (CIs). Secondary tests, including those on held-out DaLiA subjects, are reported without correction. Because the device clocks agree only to a few seconds, the primary results were repeated with label lags of -4 to $+4$ s. The PPG-Sport architectures did not reach their published error when we retrained them on DaLiA (Section VI), and PPG-Sport results are therefore paired comparisons in our pipeline.

C. Metrics

The main metric is the mean absolute error (MAE) in bpm, averaged per subject and then over subjects, as in the DaLiA benchmark [1]. To relate it to a standard, we report the median over subjects of the mean absolute percentage error (MAPE). A median below 10% is the accuracy criterion that Charlton et al. derived from the ANSI/AAMI and CTA-2065 standards for heart rate monitors [20]. A constant guess can meet such a criterion when HR varies little, and we add three checks. ACC10 is the share of windows within 10% or 5 bpm of the reference. The skill score $1 - \text{MAE}/\text{MAE}_c$ compares the MAE with that of a constant at the median label of the pool, 141 bpm. The area under the receiver operating characteristic curve (AUC) of the estimate for windows above 140 bpm measures ranking independently of calibration.

V. RESULTS

A. Main Result

Table II compares RCDL with the no-pool control. RCDL reduced the error on all five external datasets, by 19% to 40%,

TABLE II
MAE (BPM) OF RCDL AND THE NO-POOL CONTROL

Dataset	No pool	RCDL	Gain [95% CI]	p	Improved
<i>Development datasets</i>					
ScientISST	17.6	14.2	+3.4 [+0.9, +6.0]	0.033	12/16
GalaxyPPG	23.1	14.1	+9.0 [+6.6, +11.7]	<0.001	18/18
PPG-Sport	19.4	12.6	+6.8 [+4.2, +9.5]	<0.001	24/30
<i>Held-out dataset and smartwatch</i>					
WEEE	18.0	11.9	+6.1 [+2.0, +11.2]	0.033	11/14
Galaxy Watch	32.8	19.8	+13.0 [+8.5, +17.9]	<0.001	16/17
Held-out DaLiA	4.38	4.34	+0.04 [−0.06, +0.16]	0.56	9/15

Gain: no pool minus RCDL, with a subject-level bootstrap CI. p : paired Wilcoxon signed-rank test, Holm-corrected over the five external datasets (the DaLiA test is outside this family). Improved: subjects with a lower error with RCDL. Bold: $p < 0.05$ and a CI above zero.

and improved most subjects on each. On the held-out WEEE dataset, the error fell from 18.0 to 11.9 bpm. The gains on WEEE and the Galaxy Watch held at label lags of ± 2 and ± 4 s. The watch gain also held under the other PPG polarity (12.9 bpm, $p = 0.002$) and with a 0.5–8 Hz band (8.7 bpm, $p < 0.001$), under which RCDL reached 14.5 bpm. On held-out DaLiA subjects, the error of RCDL was 0.04 bpm lower (95% CI of the gain, −0.06 to +0.16 bpm, $p = 0.56$).

B. Ablations

1) *Pipeline Components*: Table III changes one component at a time, and each comparison below differs in one factor unless stated (paired tests, Holm-corrected over the five external datasets). Replacing the head with the scalar output of PULSE, with the tracker removed from both arms, raised the error by 3.4 to 17.3 bpm on all five datasets (Holm $p \leq 0.043$). This comparison also differs in the training budget (20 epochs instead of ten). With the scalar output, the pool still reduced the error of the student by 3.1 to 13.5 bpm on all five datasets (Holm $p \leq 0.022$), and the pool and the head thus each reduce the error. The tracker reduced the error on four datasets but significantly only on PPG-Sport and on held-out DaLiA subjects (0.44 bpm, $p < 0.001$). With a scalar output, the larger PULSE teacher had a lower error than the student on four datasets, significantly only on PPG-Sport (7.0 bpm, Holm $p < 0.001$) and on held-out DaLiA subjects (0.50 bpm, $p < 0.001$). For the teacher, the pool raised the DaLiA error from 3.85 to 4.06 bpm.

2) *Pool Content, Share, and Selection Rule*: We tested these choices with the scalar PULSE teacher (Table IV). Seated E4 data gave no gain, and re-weighting DaLiA’s own windows above 130 bpm recovered less than a quarter of the gain. A control with the speed-up factors of a synthetic pool, but with labels at or below 120 bpm, changed the error by at most 2.0 bpm. The gain of the synthetic pool comes from the covered range: it beat this control on the three E4 datasets by 5.0 to 8.4 bpm (Holm $p \leq 0.007$), although the control also draws lower-HR DaLiA spans. The mixed pool was the most accurate on the three E4 datasets. The real pool had a slightly lower mean error on the development datasets (19.3

TABLE III
MAE (BPM) WHEN ONE COMPONENT OF RCDL IS CHANGED

Variant	E4			Other		
	SciST	Galaxy	WEEE	Sport	Watch	DaLiA
RCDL (full)	14.2	14.1	11.9	12.6	19.8	4.34
w/o pool	17.6	23.1	18.0	19.4	32.8	4.38
w/o tracker	14.7	14.0	13.1	14.0	22.2	4.78
scalar output	18.1	21.5	18.0	27.3	39.6	4.56
scalar output, w/o pool	22.0	35.0	21.2	30.8	46.6	4.55
<i>PULSE teacher, scalar output</i>						
with pool	16.7	20.8	19.5	20.3	35.4	4.06
w/o pool	22.6	36.2	22.4	28.7	50.1	3.85

SciST = ScientISST, Galaxy = GalaxyPPG, Sport = PPG-Sport, Watch = Galaxy Watch, DaLiA = held-out DaLiA subjects. Scalar output: PULSE’s regression output, ten epochs, untracked. Bold: lowest in the column.

TABLE IV
MAE (BPM) OF THE SCALAR PULSE TEACHER WITH DIFFERENT POOLS, SHARES, AND SELECTION RULES

Pool	E4			Other	
	SciST	Galaxy	WEEE	Sport	Watch
Real pool, share $\approx 26\%$ (default)	16.7	20.8	19.5	20.3	35.4
<i>Pool content</i>					
Seated E4 data (GalaxyPPG)	22.8	–	22.6	28.6	–
Re-weighted DaLiA, >130 bpm	21.3	33.7	21.8	26.7	49.0
Stretch only, labels ≤ 120 bpm	22.7	34.4	20.4	30.2	50.3
Synthetic (sped-up DaLiA)	16.1	26.0	15.4	34.7	54.7
Half real, half synthetic	14.6	19.2	15.2	25.5	47.8
<i>Pool share</i>					
Share $\approx 3.4\%$ (no oversampling)	18.3	25.4	20.2	23.9	–
Share $\approx 10\%$	17.2	21.8	19.6	22.6	39.6
Share $\approx 52\%$	17.2	22.4	20.1	19.3	35.4
<i>Model selection and control</i>					
Early stopping on DaLiA	19.4	30.8	22.6	26.1	–
No pool (DaLiA only)	22.6	36.2	22.4	28.7	50.1

Abbreviations as in Table III. Fixed ten epochs unless stated, MAE averaged over two seeds. Watch: 0.5–4 Hz PPG band. –: not evaluated. Bold: lowest in the column.

against 19.7 bpm), because the mixed pool lost on PPG-Sport. On the two non-E4 wristbands, the synthetic pool increased the error, while the mixed pool reduced it by only 3.2 and 2.3 bpm, against 8.4 and 14.7 bpm for the real pool. Without oversampling ($r = 1$, a share of 3.2%–3.6%), the pool already reduced the error on all four datasets tested. Early stopping on DaLiA validation subjects, with at most 20 epochs instead of a fixed ten, returned the pretrained model in 20 of 30 runs. It raised the error on all four datasets (Holm $p \leq 0.031$).

C. Where the Gain Comes From

Fig. 3 shows the mean estimate per reference-HR bin. Without the pool, the estimates on the E4 datasets fall increasingly below the identity above about 140 bpm, consistent with (1). On GalaxyPPG, they even decrease above about 165 bpm, which (1) alone does not predict. With the pool, the estimates follow the reference further. Table V gives the error per HR band on the evaluated windows. Above 140 bpm, the pool reduced the error by 5.9 to 13.0 bpm on the three E4 datasets

TABLE V
MAE (BPM) PER REFERENCE HR BAND OF RCDL AND THE NO-POOL CONTROL

Dataset	100–140 bpm			>140 bpm			Windows >140
	No	RCDL	Gain	No	RCDL	Gain	
ScientISST	12.4	9.2	+3.2	18.8	12.9	+5.9	6928
GalaxyPPG	18.9	15.9	+2.9	26.4	14.0	+12.4	1196
WEEE	11.8	9.5	+2.3	23.4	15.3	+8.2	1207
PPG-Sport	18.7	10.3	+8.4	20.3	15.6	+4.8	6276
Galaxy Watch	23.3	11.6	+11.7	37.7	24.6	+13.0	1177

Windows pooled per dataset. No: no-pool control. Gain: No minus RCDL. Below 100 bpm: omitted, fewer than 100 windows on three datasets.

TABLE VI
METRIC AUDIT OF RCDL, THE NO-POOL CONTROL, AND THE CONSTANT AT 141 BPM

Dataset	MAPE (%)		ACC10 (%)		Skill		AUC	
	RCDL	No Const.	RCDL	No	RCDL	No	RCDL	No
ScientISST	6.5	9.4	27.8	71	63	0.50	0.38	0.94
GalaxyPPG	9.9	16.4	11.8	65	44	0.20	-0.31	0.69
WEEE	6.9	9.5	15.3	73	62	0.43	0.14	0.86
PPG-Sport	3.7	12.2	6.3	75	58	-0.18	-0.82	0.77
Galaxy Watch	9.2	20.8	11.3	60	31	-0.16	-0.92	0.68

No: no-pool control. Const.: constant at 141 bpm. MAPE: median over subjects, bold when below the 10% criterion. Skill: $1 - \text{MAE}/\text{MAE}_c$ from the subject-averaged MAEs, negative when worse than the constant.

and the watch. Between 100 and 140 bpm, the gain was larger on PPG-Sport (8.4 against 4.8 bpm above 140 bpm) and nearly as large on the watch (11.7 against 13.0 bpm). The gain is thus not confined to the uncovered range. In Fig. 3, the PPG-Sport windows between 150 and 175 bpm come mostly from racket and ball sports, which the running evaluation excludes, and both models fail there.

D. Metric Audit

Table VI reports the checks of Section IV-C. RCDL met the 10% MAPE criterion on all five datasets, the no-pool control on two, and the constant at 141 bpm only on PPG-Sport. There, the running HR of a subject varies little (within-subject standard deviation 5.0 bpm, against 16.4 to 28.4 bpm on the other datasets). The criterion can also hide failures. On the watch, RCDL met it and the constant did not, although the constant had the lower MAE (17.1 against 19.8 bpm). The median ignores the five watch subjects on whom RCDL erred by more than 30 bpm. The skill score of RCDL was negative on PPG-Sport and, with the planned 0.5–4 Hz band, on the watch. All metrics in Table VI favored RCDL on every dataset, as did the root-mean-square error and the concordance correlation (not shown). MAPE and ACC10 improved significantly on every dataset ($p \leq 0.039$, uncorrected), and the AUC was higher with the pool on all five datasets, with a bootstrap CI above zero on three.

TABLE VII
MAE (BPM) OF PUBLISHED MODELS, CONSTANT PREDICTORS, AND RCDL

Method	Params	SciST	Galaxy	WEEE	Sport	Watch
Constant, DaLiA median	–	53.5	61.2	50.1	55.2	61.8
Constant, pool median	–	28.3	17.7	21.0	10.7	17.1
Constant, test median ^a	–	28.0	16.8	19.7	10.2	16.2
BeliefPPG [5]	138k	10.9	16.9	7.9	7.9	–
KID-PPG [4]	378k	12.4	12.1	15.7	–	–
Pulse-PPG [11]	28.5M	23.9	42.1	–	31.7	–
Teacher, scalar, pool	132k	16.7	20.8	19.5	20.3	35.4
RCDL (ours)	76k	14.2	14.1	11.9	12.6	19.8

Constants: DaLiA median 85 bpm, pool median 141 bpm. ^aOracle (test labels). Abbreviations as in Table III. BeliefPPG and KID-PPG: released weights, KID-PPG decoded with a 4-s smoother. Pulse-PPG: fine-tuned on DaLiA. Params: as published for BeliefPPG and Pulse-PPG, counted otherwise. –: not evaluated. Bold: lowest, excluding the oracle.

E. Comparison With Published Models and Constants

In Table VII, RCDL was more accurate than the scalar teacher pipeline on all five datasets, significantly on four (Holm $p \leq 0.037$). Pulse-PPG, fine-tuned on DaLiA, was the least accurate trained model on every dataset it was tested on. BeliefPPG was the most accurate model on three of the four datasets on which it was evaluated, and KID-PPG, decoded with a 4-s smoother, was more accurate than RCDL on ScientISST and GalaxyPPG. In uncorrected paired tests on the same windows, RCDL did not differ significantly from BeliefPPG or KID-PPG on GalaxyPPG ($p \geq 0.18$) or from KID-PPG on ScientISST ($p = 0.27$). It was less accurate than BeliefPPG on ScientISST ($p = 0.034$). RCDL has about half the parameters of BeliefPPG and a fifth of those of KID-PPG.

VI. LIMITATIONS

RCDL was selected on the development datasets, the rule of its first pre-registered test on WEEE was not met, and the watch shares participants, sessions, and labels with GalaxyPPG. The real pool adds running motion as well as high heart rates, as the gains between 100 and 140 bpm suggest, and we did not test a running pool without high heart rates. The head adds a spectral input and a distributional output, and our ablation does not separate the two. We compared the pool content, its share, and the selection rule with the scalar teacher only, and neither the number of epochs nor selection on held-out pool recordings was tested. In a separate experiment, the same head and pool raised the error of the PULSE teacher on held-out DaLiA subjects by 2.0 bpm (5.79 against 3.83 bpm for its smoothed scalar version with the pool). BeliefPPG remains more accurate on most datasets, and RCDL still underestimates the highest heart rates (Fig. 3). We could not adapt BeliefPPG or KID-PPG: fine-tuning KID-PPG without its original augmentation degraded it, and retraining BeliefPPG from its released code diverged in our runs. For the scalar teacher on the watch, a 0.5–8 Hz band made the synthetic pool the best option (22.5 against 24.2 bpm for the real pool, exploratory). The reference HR of the E4 datasets

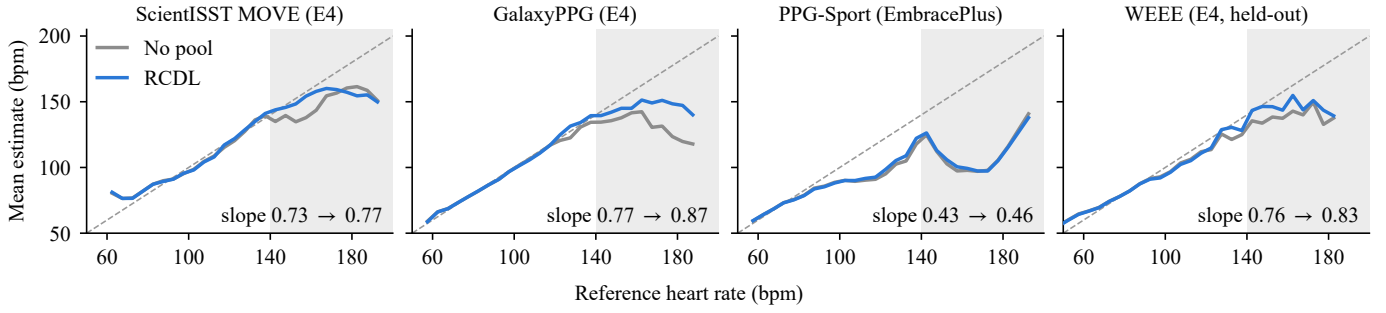


Fig. 3. Mean estimate per 5-bpm bin of reference HR over all valid windows of four datasets, for RCDL and the no-pool control. The dashed line is the identity, the shaded range is the uncovered range, and the slopes are least-squares fits of the estimate on the reference.

is our own ECG consensus, so absolute errors are not directly comparable with published results on these datasets. Trained on DaLiA's periodic activities as in the dataset paper, the two PPG-Sport architectures reached at best 19.4 and 18.2 bpm on its running data, against the published 8.7 bpm. Finally, each configuration was trained with two seeds, and the pool contains 12 recordings.

VII. CONCLUSION

DaLiA-trained PPG heart rate estimators underestimate high heart rates during running, in part because their training data do not cover them. Range-Covering Deep Learning fine-tunes a small pretrained encoder with a spectral distribution head, a small pool of real treadmill recordings, a fixed training budget, and a shift-invariant tracker. Chosen on three development datasets, it reduced the error on all five external datasets, including a dataset held out until the pipeline was fixed, without a detectable change on DaLiA. Separating the roles of the HR range and of the running motion in the pool is the next step. We will release the pre-registered scripts and results upon acceptance.

REFERENCES

- [1] A. Reiss, I. Indlekofer, P. Schmidt, and K. Van Laerhoven, "Deep PPG: Large-scale heart rate estimation with convolutional neural networks," *Sensors*, vol. 19, no. 14, p. 3079, 2019.
- [2] P. Kasnesis, L. Toumanidis, A. Burrello, C. Chatzigeorgiou, and C. Z. Patrikakis, "Feature-level cross-attentional PPG and motion signal fusion for heart rate estimation," in *Proc. IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, 2023, pp. 1731–1736.
- [3] P. Kasnesis, L. Toumanidis, D. Jahier Pagliari, and A. Burrello, "Replacing attention with modality-wise convolution for energy-efficient PPG-based heart rate estimation using knowledge distillation," *IEEE J. Biomed. Health Inform.*, vol. 30, no. 1, pp. 339–352, 2026.
- [4] C. Kechris, J. Dan, J. Miranda, and D. Atienza, "KID-PPG: Knowledge informed deep learning for extracting heart rate from a smartwatch," *IEEE Trans. Biomed. Eng.*, vol. 72, no. 3, pp. 870–877, 2025.
- [5] V. Bieri, P. Strel, B. U. Demirel, and C. Holz, "BeliefPPG: Uncertainty-aware heart rate estimation from PPG signals via belief propagation," in *Proc. 39th Conference on Uncertainty in Artificial Intelligence (UAI)*, ser. PMLR, vol. 216, 2023, pp. 173–183.
- [6] S. Park, D. Zheng, and U. Lee, "A PPG signal dataset collected in semi-naturalistic settings using Galaxy Watch," *Sci. Data*, vol. 12, p. 892, 2025.
- [7] Y. Yang, K. Zha, Y. Chen, H. Wang, and D. Katabi, "Delving into deep imbalanced regression," in *Proc. 38th International Conference on Machine Learning (ICML)*, ser. PMLR, vol. 139, 2021, pp. 11 842–11 851.
- [8] M. Risso *et al.*, "Robust and energy-efficient PPG-based heart-rate monitoring," in *Proc. IEEE International Symposium on Circuits and Systems (ISCAS)*, 2021, pp. 1–5.
- [9] A. Burrello *et al.*, "Improving PPG-based heart-rate monitoring with synthetically generated data," in *Proc. IEEE Biomedical Circuits and Systems Conference (BioCAS)*, 2022, pp. 153–157.
- [10] L. Benfenati *et al.*, "EnhancePPG: Improving PPG-based heart rate estimation with self-supervision and augmentation," in *Proc. IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, 2025, pp. 1–5.
- [11] M. Saha, M. A. Xu, W. Mao, S. Neupane, J. M. Rehg, and S. Kumar, "Pulse-PPG: An open-source field-trained PPG foundation model for wearable applications across lab and field settings," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 9, no. 3, pp. 1–35, 2025.
- [12] L. Torgo, R. P. Ribeiro, B. Pfahringer, and P. Branco, "SMOTE for regression," in *Proc. Portuguese Conference on Artificial Intelligence (EPIA)*, ser. LNCS, vol. 8154, 2013, pp. 378–389.
- [13] H. Yao, Y. Wang, L. Zhang, J. Zou, and C. Finn, "C-Mixup: Improving generalization in regression," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, 2022, pp. 3361–3376.
- [14] Z. Lipton, Y.-X. Wang, and A. Smola, "Detecting and correcting for label shift with black box predictors," in *Proc. 35th International Conference on Machine Learning (ICML)*, ser. PMLR, vol. 80, 2018, pp. 3122–3130.
- [15] Z. Zhang, Z. Pi, and B. Liu, "TROIKA: A general framework for heart rate monitoring using wrist-type photoplethysmographic signals during intensive physical exercise," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 2, pp. 522–531, 2015.
- [16] J. Areias Saraiva, M. Abreu, A. S. Carmo, H. Plácido da Silva, and A. Fred, "ScientISST MOVE: Annotated wearable multimodal biosignals recorded during everyday life activities in naturalistic environments (version 1.0.1)," *PhysioNet*, 2024. [Online]. Available: <https://doi.org/10.13026/hyxq-r919>
- [17] C. Hu *et al.*, "PPG-Sport: A dataset for reliable heart rate monitoring from wrist PPG under dynamic sports conditions," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 10, no. 2, pp. 1–28, 2026.
- [18] S. Gashi, C. Min, A. Montanari, S. Santini, and F. Kawsar, "A multidevice and multimodal dataset for human energy expenditure estimation using wearable devices," *Sci. Data*, vol. 9, p. 537, 2022.
- [19] D. Makowski *et al.*, "NeuroKit2: A Python toolbox for neurophysiological signal processing," *Behav. Res. Methods*, vol. 53, no. 4, pp. 1689–1696, 2021.
- [20] P. H. Charlton *et al.*, "Detecting beats in the photoplethysmogram: Benchmarking open-source algorithms," *Physiol. Meas.*, vol. 43, no. 8, p. 085007, 2022.