RELATIONAL WINDOW STABILIZATION FOR REPEATABLE SINGLE-WINDOW PPG DISEASE-BACKGROUND PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep learning models for physiological signals are increasingly used for continuous monitoring of chronic disease background, yet their evaluation mainly relies on population-level discrimination metrics such as AUROC and AUPRC. In continuous monitoring, however, repeated measurements from the same individual should produce repeatable predictions when the underlying disease-background state remains unchanged over short intervals. Recent studies have revealed substantial prediction variability across repeated physiological measurements, highlighting a gap between discrimination performance and prediction repeatability.

Here, we introduce Relational Window Stabilization, a training framework that learns repeatable single-window PPG disease-background predictions by exploiting relationships among multiple windows collected within the same monitoring episode. RWS uses multiple windows only during training, preserving individual-window supervision while reducing within-episode prediction variation without requiring inference-time aggregation.

We evaluate RWS on the MC-MED PPG benchmark across nine chronic ICD-derived disease-background tasks using patient-disjoint chronological splits. Across different input durations, architectures, representations, and random seeds, RWS improves both discrimination and repeatability, increasing AUROC, AUPRC, and intraclass correlation coefficient while reducing prediction variability and diagnostic flip rates. These improvements remain under matched multi-window exposure controls and alternative repeatability-learning strategies. Cross-dataset evaluation on MIMIC-III further demonstrates that the repeatability advantage persists after checkpoint transfer and independent in-domain retraining.

Together, these findings establish prediction repeatability as a complementary evaluation dimension for continuous physiological AI and demonstrate that same-episode relational supervision can improve single-window PPG predictions without changing the deployment paradigm.

1 INTRODUCTION

Deep learning models have achieved remarkable success in learning predictive representations from complex physiological signals and have enabled automated assessment of diverse disease-background states. Existing studies have demonstrated that models based on signals such as electrocardiography (ECG) and photoplethysmography (PPG) can extract clinically relevant information beyond conventional measurements, supporting applications including cardiac abnormality detection and continuous physiological monitoring (Attia et al., 2019a;b; Hannun et al., 2019; Tison et al., 2018; Guo et al., 2021).

However, most current models are optimized and evaluated according to population-level discrimination metrics, such as AUROC and AUPRC. These metrics measure whether predictions can distinguish individuals with different disease-background states, but do not characterize whether a model produces repeatable predictions when repeatedly observing the same individual. This distinction is becoming increasingly important as physiological sensing shifts from intermittent measurements toward continuous monitoring. In continuous monitoring scenarios, a single individual can generate a

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

large number of signal windows over time, each of which may be independently evaluated by an AI model. For chronic disease-background states or other phenotypes expected to remain unchanged over short intervals, the underlying clinical state should not substantially change within minutes. Therefore, repeated observations from the same individual would be expected to produce similar model outputs.

Nevertheless, physiological signals naturally vary across repeated windows because of differences in waveform morphology, local temporal patterns, and measurement conditions. Models trained with conventional supervised objectives may therefore produce different predictions across repeated observations, even when the underlying state remains unchanged. Such prediction variability represents a distinct limitation from poor discrimination performance: a model may accurately separate populations while still lacking repeatability for repeated monitoring of the same individual.

Recent clinical AI studies have provided direct evidence of this phenomenon. Hamada et al. evaluated an AI-based atrial-fibrillation risk-estimation model using repeated ECG recordings and showed that prediction agreement decreased as the interval between recordings increased: the quadratic weighted κ was 0.72 for ECGs repeated within four minutes but decreased to 0.47 for recordings separated by approximately 15 minutes (Hamada et al., 2026). This observation highlights that model predictions can be sensitive to input variation even when the underlying clinical state is expected to remain unchanged over short intervals, suggesting that discrimination performance and prediction repeatability represent different properties of learned models.

Despite its importance, prediction repeatability has not been systematically studied as a learning objective. Existing approaches for physiological-signal models mainly focus on improving representation quality, increasing discriminative performance, or reducing input-level measurement variability. These approaches do not explicitly optimize whether repeated observations from the same individual should produce repeatable predictions. This raises an important machine-learning question:

Can repeated observations of the same individual be leveraged during training to improve the repeatability of single-observation predictions while preserving discriminative ability?

In this work, we study this question in the context of continuous PPG-based disease-background prediction and introduce Relational Window Stabilization (RWS), a training framework that learns repeatable predictions from relationships among multiple windows collected within the same monitoring episode. RWS exploits repeated windows as relational information during training, encouraging repeatable outputs among observations expected to share the same underlying state while preserving discrimination across different episodes. Importantly, repeated windows are used only during training, while inference remains based on a single signal window.

We evaluate RWS across diverse PPG modeling settings, including different input durations, architectures, and representation strategies. Beyond in-domain evaluation, we investigate whether the learned repeatability property transfers across datasets through frozen-checkpoint transfer and independent training on MIMIC-III-Ext-PPG. Our results demonstrate that prediction repeatability can be improved as a property of the learned prediction function rather than through inference-time aggregation or model-specific architectural modifications.

1.1 Related work and scope

Deep learning for physiological-signal prediction. Deep learning models have been extensively applied to physiological signals for automated disease recognition and health-state assessment. Recent studies and benchmarks have explored different architectures, input durations, and representation-learning strategies to improve predictive performance. For example, SignalMC-MED provides a standardized evaluation framework for comparing ECG- and PPG-based models across multiple disease-recognition tasks (Gustafsson et al., 2026). However, these evaluations primarily focus on population-level discrimination metrics and do not assess whether predictions remain repeatable across naturally repeated observations from the same individual.

Prediction repeatability of physiological AI models. Compared with discrimination performance, the repeatability of model outputs under repeated observations has received limited attention. Previous studies have investigated variability arising from repeated physiological measurements through signal processing, denoising, and measurement-variation analysis (McLaughlin et al., 1994;

1996). More recently, AI-based ECG studies have directly quantified prediction variability across repeated recordings and demonstrated that model outputs can change over short intervals even when the underlying clinical state is expected to remain unchanged (Hamada et al., 2026).

However, existing studies have primarily characterized prediction variability or addressed input-level measurement variation rather than optimizing repeatability during model training. In particular, whether naturally repeated observations can be incorporated as relational supervision to learn repeatable predictors remains unclear. Continuous physiological monitoring provides a unique opportunity for studying this problem because repeated signal windows from the same individual are naturally available, while practical deployment often requires predictions from individual observations.

In this work, we formulate short-term prediction repeatability as an additional property of predictive models under repeated observations and propose RWS to optimize this property. During training, RWS leverages same-episode multi-window relationships to constrain within-episode prediction variation while preserving individual-window supervision and single-window inference.

Our contributions are summarized as follows:

- We identify short-term prediction repeatability as an overlooked property of deep learning models under repeated observations, and demonstrate that conventional discrimination metrics do not capture within-individual prediction variability.
- We propose RWS, a general training framework that exploits same-episode relational information to improve single-window prediction repeatability while maintaining discriminative supervision and the original inference paradigm.
- We systematically evaluate repeatability learning across diverse architectures, input durations, representation strategies, and datasets, demonstrating that the learned repeatability improvement remains robust under different modeling conditions and dataset shifts.

2 METHOD

2.1 PROBLEM FORMULATION

For monitoring episode v, let $\{x_{v,k}\}_{k=1}^{K}$ be its waveform-window bank and let $y_v \in \{0,1\}^T$ be the task label vector. A shared encoder and task head produce one logit and probability vector per window,

$$h_{v,k} = f_\theta(x_{v,k}), \qquad z_{v,k} = g_\phi(h_{v,k}), \qquad p_{v,k} = \sigma(z_{v,k}). \tag{1}$$

RWS uses the bank during training to learn a single-window predictor: each window remains task-predictive, while outputs from the same episode are constrained to vary less than outputs between episodes. The relational terms are training-only; the primary deployment protocol uses one new window.

2.2 SAME-EPISODE RELATIONAL OBJECTIVE

RWS combines task supervision with three relational terms,

$$\mathcal{L}_{\text{total}}(e) = \lambda_{\text{task}}\mathcal{L}_{\text{task}} + r(e)\left(\lambda_{\text{cons}}\mathcal{L}_{\text{cons}} + \lambda_{\text{range}}\mathcal{L}_{\text{range}} + \lambda_{\text{ICC}}\mathcal{L}_{\text{ICC}}\right), \tag{2}$$

where $r(e)$ warms up the relational terms during training. The task term is weighted BCE-with-logits applied to every window,

$$\mathcal{L}_{\text{task}} = \frac{1}{KT}\sum_{k,t}\text{BCEWithLogits}(z_{v,k,t}, y_{v,t}; w_t^+). \tag{3}$$

The leave-one-out logit target couples windows without including the current prediction in its own target,

$$\widetilde{z}_{v,k,t} = \text{stopgrad}\left(\frac{1}{K-1}\sum_{\substack{j=1\\ j\neq k}}^{K} z_{v,j,t}\right), \qquad \mathcal{L}_{\text{cons}} = \frac{1}{KT}\sum_{k,t}\text{SmoothL1}(z_{v,k,t}, \widetilde{z}_{v,k,t}). \tag{4}$$

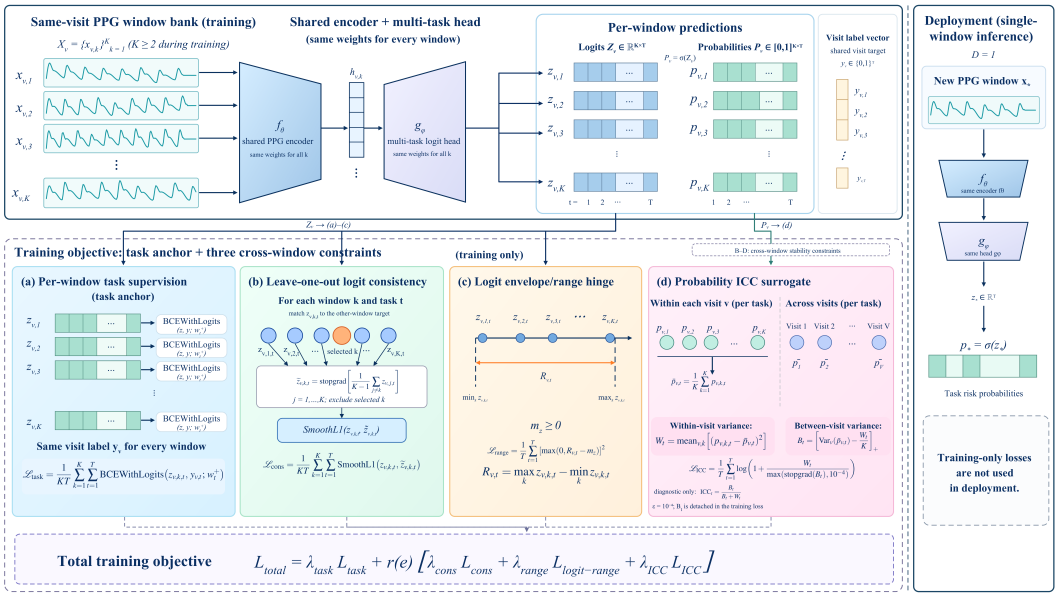


Figure 1: RWS mechanism.

The logit-envelope term limits excessive within-episode disagreement,

$$R_{v,t} = \max_k z_{v,k,t} - \min_k z_{v,k,t}, \qquad \mathcal{L}_{\text{range}} = \frac{1}{T}\sum_t \left[\max(0, R_{v,t} - m_z)\right]^2 . \tag{5}$$

To retain between-episode separation, let W_t be within-episode probability variance and let B_t be the finite-window-corrected variance of episode means:

$$W_t = \frac{1}{VK}\sum_{v,k}(p_{v,k,t} - \bar{p}_{v,t})^2, \qquad B_t = \left[\mathrm{Var}_v(\bar{p}_{v,t}) - \frac{W_t}{K}\right]_+ , \tag{6}$$

where $\bar{p}_{v,t} = K^{-1}\sum_k p_{v,k,t}$. The training surrogate is

$$\mathcal{L}_{\text{ICC}} = \frac{1}{T}\sum_t \log\left(1 + \frac{W_t}{\max(\mathrm{stopgrad}(B_t), 10^{-4})}\right). \tag{7}$$

The detached denominator prevents the objective from being reduced by merely spreading episode means apart.

3 Experimental Setup

3.1 Dataset, Tasks, and Split

We use the MC-MED PPG Pleth benchmark with nine chronic ICD-derived disease-background tasks; Supplementary Table S1 gives the duration-specific cohort and task distribution. Demographic features are excluded. The task set and representative public pretrained encoders follow the SignalMC-MED benchmark protocol (Gustafsson et al., 2026); the source MC-MED cohort is described by Kansal et al. (Kansal et al., 2025). Labels are multi-label and are macro-averaged with equal task weight. Because these chronic disease-background labels are not acute outcomes expected to change during one acquisition, windows from one episode are treated as exchangeable views of the same short-term label state. We use a patient-disjoint chronological split with no shared MRN or CSN across partitions.

The primary comparison uses seeds 42–44, $K = 8$ training windows, $E = 8$ evaluation windows, and $D = 1$ deployment window. Paired conditions use the same split and evaluation bank. The

separate inference-only audit orders feasible starts chronologically and places eight starts at equal intervals from the earliest to latest start within each episode. Each model forward still scores one window. The mean temporal spacing between adjacent selected starts is 1092.9 s for 10-s inputs, 1088.4 s for 30-s inputs, and 1090.7 s for 60-s inputs. For deployment count D, the ordered bank is partitioned into consecutive non-overlapping groups of D windows and the group logits are averaged: $D = 1$, $D = 2$, and $D = 4$ yield 8, 4, and 2 final predictions per episode, respectively, whereas $D = 8$ yields one final prediction. The checkpoint and bank are unchanged; only the number of windows contributing to each final prediction is varied.

3.2 Cross-dataset transfer and independent replication

We additionally evaluate the same nine ICD-derived disease-background tasks on the full MIMIC-III-Ext-PPG cohort (6,780 HADM records from 6,189 subjects), whose data resource is described by Moulaeifard et al. (Moulaeifard et al., 2026).We use one monitoring episode as the unit of repeated observation: in MC-MED, an episode corresponds to one visit/CSN, whereas in MIMIC-III-Ext-PPG it corresponds to one HADM_ID.For MIMIC-III-Ext-PPG, the nine task labels are joined at HADM_ID using the audited ICD-9 crosswalk so that the target-domain task definitions match the MC-MED task specifications.We use the official v1.1.0 subject-disjoint assignments; all records from one subject remain in a single fold and that no subject appears in more than one of the ten folds. The PPG is represented at 125 Hz and each input window is independently z-scored, matching the active MC-MED signal preprocessing.The MIMIC release provides fixed 30-s records.10-s inputs are the 0-, 10-, or 20-s slice of one 30-s record, 30-s inputs use one complete record, and 60-s inputs concatenate two records whose timestamps differ by exactly 30 s. Thus a 60-s input never crosses an acquisition gap or HADM. Eight candidate windows are then selected by equal temporal coverage.

For checkpoint transfer, the MC-MED encoder, head, and checkpoint are frozen. AUROC, AUPRC, ICC, and within-episode SD are computed from the native transferred probabilities without target-domain model adaptation. The primary flip rate and value-dependent flip rate (VDFR) use 5-fold cross-fitted target-domain Platt calibration and held-out Youden thresholds; each held-out HADM is excluded from its calibrator. A source-frozen calibration sensitivity uses no MIMIC labels and tests whether the decision-level repeatability direction is preserved without target-domain calibration. Independent in-domain replication calibrates on fold 8, and tests on fold 9, separating transfer from learning the same principle in the new dataset. Holm and BH–FDR are applied separately to the 18-test checkpoint-transfer family and the 18-test in-domain-retraining family.

3.3 Architectures and comparison conditions

The cross-architecture matrix contains three end-to-end models (BiGRU, ResNet1D, and PatchTST) and four frozen-encoder representations (CSFM, PaPaGei, MOMENT, and ECGFounder) with a unified MLP head (Gu et al., 2026; Pillai et al., 2025; Goswami et al., 2024; Li et al., 2025). ECGFounder is an ECG-pretrained representation transferred to PPG. Each architecture–duration–seed cell compares Conventional single-window BCE ($K = 1$) with RWS trained on $K = 8$ windows under the same split, seed, evaluation bank, and $D = 1$ setting.

The fair-exposure matrix compares Conventional BCE ($K = 1$), BCE-only multi-window training ($K = 8$), and RWS ($K = 8$). The matched BiGRU method-comparison matrix contains self-relation, teacher–student, MIL top-2, standard multi-view, and EMA mean-teacher controls (Tarvainen & Valpola, 2017); the ablation matrix removes one RWS relational component at a time.

3.4 Metrics and uncertainty

The six primary endpoints are AUROC, AUPRC, ICC, within-episode SD (WSD), diagnostic flip rate, and VDFR. AUROC, AUPRC, and ICC are higher-is-better; WSD, flip rate, and VDFR are lower-is-better. Main tables report absolute three-seed mean ± SD.

Let $p_{v,k,t}$ be the native sigmoid probability and $\bar{p}_{v,t} = K^{-1} \sum_k p_{v,k,t}$. The absolute within-episode fluctuation is

$$\mathrm{WSD} = \frac{1}{VT} \sum_{v,t} \left[\frac{1}{K} \sum_{k=1}^{K} (p_{v,k,t} - \bar{p}_{v,t})^2 \right]^{1/2}. \tag{8}$$

For each task, the $E = 8$ same-episode windows are treated as repeated ratings. Let $\bar{p}_v = E^{-1}\sum_e p_{v,e}$ and $\bar{p} = V^{-1}\sum_v \bar{p}_v$. The standard one-way random-effects, single-measure ICC is computed from

$$MS_{\text{between}} = \frac{E}{V-1}\sum_v (\bar{p}_v - \bar{p})^2, \qquad MS_{\text{within}} = \frac{1}{V(E-1)}\sum_{v,e}(p_{v,e} - \bar{p}_v)^2. \tag{9}$$

$$\text{ICC} = \frac{MS_{\text{between}} - MS_{\text{within}}}{MS_{\text{between}} + (E-1)MS_{\text{within}}}. \tag{10}$$

The reported ICC is macro-averaged over tasks. Diagnostic flip uses validation-calibrated probabilities and the validation-Youden threshold τ_t:

$$\text{Flip}_{v,t} = \mathbb{1}\left[\min_k p^{\text{cal}}_{v,k,t} < \tau_t \le \max_k p^{\text{cal}}_{v,k,t}\right]. \tag{11}$$

We call the resulting diagnostic measure the VDFR. It is the fraction of the fixed grid $\mathcal{T} = \{0.05, 0.10, \ldots, 0.95\}$ crossed by the same-episode probability interval:

$$\text{VDFR}_{v,t} = \frac{1}{|\mathcal{T}|}\sum_{\tau\in\mathcal{T}} \mathbb{1}\left[\min_k p^{\text{cal}}_{v,k,t} < \tau \le \max_k p^{\text{cal}}_{v,k,t}\right]. \tag{12}$$

WSD and ICC use native probabilities; flip rate and VDFR use validation-calibrated probabilities. For probability-quality diagnostics, a validation-fitted Platt calibrator is applied once to held-out test logits. Supplementary tables report calibrated Brier, ECE, slope, and intercept.

Calibrated Brier and ECE compare complete validation-calibrated pipelines; slope and intercept are descriptive diagnostics, and calibration is not retraining.

For MC-MED, participant/MRN-level paired bootstrap intervals quantify test-sample uncertainty separately from the three-seed SD; Tables S8a–S8c report direction-normalized effects and S9a–S9b report multiplicity sensitivity. For MIMIC-III-Ext-PPG, 2,000 subject-level paired resamples retain all HADM records and eight windows within subject clusters and use identical draws for Conventional and RWS; these intervals are reported separately in Tables S13–S14.

3.5 Resource and inference analyses

The $K = 1$–10 scan fixes $E = 8$ and $D = 1$; the deployment audit uses completed checkpoints. Figures S1, S2, and S3 report the window-count trajectory, training resources, and deployment-window analysis, respectively.

4 Results

4.1 RWS improves repeatability across architectures

Table 1 is the primary cross-architecture result. Relative to Conventional training, RWS improves all six primary endpoints in every evaluated architecture–duration cell: AUROC, AUPRC, and ICC increase, while within-episode SD, diagnostic flip rate, and VDFR decrease. Longer inputs generally yield higher discrimination and repeatability and lower output variation.

The paired participant/MRN bootstrap gives favorable 95% intervals for all six primary endpoints throughout the cross-architecture comparison (Table S8a). All cell-level p-values in this family remain below 0.05 after both BH–FDR and Bonferroni correction (Table S9a). Task-level tables S2a–S2i show the same favorable direction for the repeatability endpoints across disease-background tasks. The calibrated counterparts are reported in Tables S3a–S3i and S4. Because BiGRU had the strongest discrimination, we used it as the representative architecture for a subgroup sensitivity audit. Each reported sex, age, and race stratum showed a favorable median comparison for all six primary endpoints; the absolute summaries are given in Tables S10–S12.

Table 1: Cross-architecture comparison of Conventional training and RWS.

		AUROC		AUPRC		ICC		Within-episode SD		Flip rate		VDFR	
Architecture	Duration	Conventional	RWS	Conventional	RWS	Conventional	RWS	Conventional	RWS	Conventional	RWS	Conventional	RWS
BiGRU	10 s	0.732 ± 0.002	**0.740 ± 0.000**	0.183 ± 0.002	**0.198 ± 0.001**	0.696 ± 0.001	**0.811 ± 0.002**	0.094 ± 0.000	**0.021 ± 0.001**	0.595 ± 0.006	**0.458 ± 0.009**	0.080 ± 0.000	**0.065 ± 0.001**
BiGRU	30 s	0.760 ± 0.001	**0.769 ± 0.001**	0.219 ± 0.002	**0.232 ± 0.001**	0.797 ± 0.002	**0.886 ± 0.001**	0.079 ± 0.001	**0.033 ± 0.001**	0.457 ± 0.007	**0.331 ± 0.004**	0.072 ± 0.000	**0.057 ± 0.001**
BiGRU	60 s	0.769 ± 0.001	**0.779 ± 0.001**	0.232 ± 0.002	**0.249 ± 0.001**	0.836 ± 0.001	**0.913 ± 0.002**	0.070 ± 0.000	**0.032 ± 0.001**	0.396 ± 0.008	**0.278 ± 0.007**	0.067 ± 0.000	**0.052 ± 0.001**
ResNet1D	10 s	0.721 ± 0.004	**0.733 ± 0.004**	0.178 ± 0.004	**0.193 ± 0.003**	0.687 ± 0.001	**0.798 ± 0.014**	0.095 ± 0.001	**0.021 ± 0.001**	0.596 ± 0.008	**0.483 ± 0.018**	0.079 ± 0.001	**0.065 ± 0.000**
ResNet1D	30 s	0.754 ± 0.001	**0.763 ± 0.001**	0.211 ± 0.001	**0.223 ± 0.001**	0.773 ± 0.002	**0.876 ± 0.001**	0.086 ± 0.000	**0.032 ± 0.001**	0.470 ± 0.014	**0.356 ± 0.001**	0.075 ± 0.001	**0.059 ± 0.000**
ResNet1D	60 s	0.762 ± 0.002	**0.773 ± 0.001**	0.222 ± 0.004	**0.235 ± 0.003**	0.814 ± 0.004	**0.902 ± 0.001**	0.077 ± 0.001	**0.032 ± 0.000**	0.417 ± 0.007	**0.304 ± 0.001**	0.070 ± 0.001	**0.054 ± 0.000**
PatchTST	10 s	0.702 ± 0.001	**0.713 ± 0.000**	0.158 ± 0.000	**0.175 ± 0.000**	0.646 ± 0.001	**0.767 ± 0.002**	0.097 ± 0.001	**0.017 ± 0.000**	0.648 ± 0.004	**0.516 ± 0.007**	0.076 ± 0.001	**0.066 ± 0.000**
PatchTST	30 s	0.734 ± 0.001	**0.741 ± 0.002**	0.186 ± 0.001	**0.200 ± 0.001**	0.741 ± 0.002	**0.854 ± 0.003**	0.090 ± 0.000	**0.025 ± 0.003**	0.533 ± 0.002	**0.386 ± 0.012**	0.072 ± 0.001	**0.058 ± 0.001**
PatchTST	60 s	0.745 ± 0.001	**0.755 ± 0.001**	0.200 ± 0.002	**0.215 ± 0.001**	0.783 ± 0.002	**0.876 ± 0.001**	0.082 ± 0.001	**0.033 ± 0.000**	0.474 ± 0.004	**0.345 ± 0.009**	0.071 ± 0.000	**0.057 ± 0.000**
CSFM	10 s	0.668 ± 0.001	**0.679 ± 0.005**	0.139 ± 0.000	**0.146 ± 0.003**	0.512 ± 0.015	**0.653 ± 0.013**	0.104 ± 0.002	**0.020 ± 0.009**	0.738 ± 0.008	**0.631 ± 0.010**	0.077 ± 0.001	**0.067 ± 0.001**
CSFM	30 s	0.698 ± 0.001	**0.706 ± 0.006**	0.156 ± 0.001	**0.167 ± 0.004**	0.623 ± 0.010	**0.764 ± 0.020**	0.101 ± 0.001	**0.019 ± 0.001**	0.645 ± 0.007	**0.488 ± 0.026**	0.075 ± 0.001	**0.062 ± 0.001**
CSFM	60 s	0.712 ± 0.001	**0.724 ± 0.002**	0.171 ± 0.001	**0.181 ± 0.001**	0.665 ± 0.012	**0.816 ± 0.010**	0.101 ± 0.003	**0.025 ± 0.002**	0.597 ± 0.014	**0.429 ± 0.020**	0.077 ± 0.001	**0.059 ± 0.001**
PaPaGei	10 s	0.622 ± 0.001	**0.635 ± 0.004**	0.107 ± 0.001	**0.114 ± 0.002**	0.381 ± 0.011	**0.532 ± 0.022**	0.099 ± 0.004	**0.001 ± 0.000**	0.841 ± 0.004	**0.687 ± 0.044**	0.073 ± 0.001	**0.063 ± 0.003**
PaPaGei	30 s	0.650 ± 0.001	**0.660 ± 0.004**	0.124 ± 0.001	**0.130 ± 0.003**	0.498 ± 0.003	**0.635 ± 0.024**	0.097 ± 0.001	**0.007 ± 0.010**	0.758 ± 0.009	**0.633 ± 0.029**	0.074 ± 0.000	**0.067 ± 0.000**
PaPaGei	60 s	0.663 ± 0.001	**0.675 ± 0.004**	0.132 ± 0.001	**0.143 ± 0.002**	0.529 ± 0.002	**0.694 ± 0.029**	0.103 ± 0.001	**0.011 ± 0.001**	0.701 ± 0.012	**0.565 ± 0.032**	0.077 ± 0.000	**0.066 ± 0.001**
MOMENT	10 s	0.641 ± 0.001	**0.658 ± 0.007**	0.116 ± 0.000	**0.127 ± 0.004**	0.444 ± 0.020	**0.588 ± 0.030**	0.107 ± 0.011	**0.004 ± 0.004**	0.809 ± 0.021	**0.698 ± 0.019**	0.075 ± 0.001	**0.069 ± 0.003**
MOMENT	30 s	0.665 ± 0.002	**0.679 ± 0.006**	0.131 ± 0.001	**0.141 ± 0.005**	0.566 ± 0.008	**0.696 ± 0.020**	0.098 ± 0.001	**0.015 ± 0.008**	0.707 ± 0.009	**0.585 ± 0.021**	0.076 ± 0.001	**0.065 ± 0.001**
MOMENT	60 s	0.681 ± 0.001	**0.694 ± 0.007**	0.141 ± 0.001	**0.154 ± 0.007**	0.632 ± 0.007	**0.767 ± 0.030**	0.091 ± 0.006	**0.023 ± 0.010**	0.645 ± 0.013	**0.498 ± 0.036**	0.075 ± 0.000	**0.063 ± 0.003**
ECGFounder	10 s	0.673 ± 0.001	**0.684 ± 0.002**	0.137 ± 0.000	**0.147 ± 0.001**	0.548 ± 0.012	**0.650 ± 0.020**	0.106 ± 0.003	**0.020 ± 0.005**	0.732 ± 0.012	**0.643 ± 0.012**	0.079 ± 0.001	**0.072 ± 0.001**
ECGFounder	30 s	0.692 ± 0.003	**0.708 ± 0.001**	0.153 ± 0.001	**0.165 ± 0.001**	0.627 ± 0.005	**0.760 ± 0.017**	0.103 ± 0.004	**0.025 ± 0.004**	0.643 ± 0.013	**0.499 ± 0.026**	0.078 ± 0.001	**0.066 ± 0.003**
ECGFounder	60 s	0.706 ± 0.001	**0.723 ± 0.002**	0.162 ± 0.000	**0.177 ± 0.001**	0.655 ± 0.016	**0.810 ± 0.003**	0.104 ± 0.003	**0.029 ± 0.000**	0.618 ± 0.014	**0.442 ± 0.000**	0.078 ± 0.001	**0.063 ± 0.001**

Each entry is an absolute three-seed mean ± SD. Conventional and RWS use the same architecture, duration, split, seed set, and evaluation bank; RWS cells are bolded. ICC is the standard one-way random-effects single-measure estimator. Paired participant/MRN bootstrap evidence is summarized in Supplementary Table S8a.

Table 2: Comparison with alternative repeatability methods on BiGRU.

Compared method	Duration	AUROC	AUPRC	ICC	Within-episode SD	Flip rate	VDFR
Self-relation	10 s	0.742 ± 0.001	0.197 ± 0.002	0.702 ± 0.006	0.099 ± 0.001	0.586 ± 0.003	0.082 ± 0.000
Teacher–student	10 s	0.740 ± 0.000	0.195 ± 0.001	0.718 ± 0.001	0.090 ± 0.001	0.568 ± 0.005	0.080 ± 0.000
MIL top-2 aggregation	10 s	0.734 ± 0.001	0.188 ± 0.001	0.672 ± 0.007	0.085 ± 0.004	0.574 ± 0.013	0.084 ± 0.001
Standard multi-view	10 s	0.744 ± 0.001	0.199 ± 0.000	0.704 ± 0.007	0.097 ± 0.001	0.576 ± 0.012	0.082 ± 0.001
EMA mean-teacher	10 s	0.744 ± 0.000	0.199 ± 0.001	0.705 ± 0.003	0.098 ± 0.001	0.579 ± 0.006	0.083 ± 0.002
RWS	**10 s**	**0.740 ± 0.000**	**0.198 ± 0.001**	**0.811 ± 0.002**	**0.021 ± 0.001**	**0.458 ± 0.009**	**0.065 ± 0.001**
Self-relation	30 s	0.764 ± 0.000	0.227 ± 0.001	0.784 ± 0.005	0.084 ± 0.002	0.459 ± 0.006	0.076 ± 0.001
Teacher–student	30 s	0.764 ± 0.000	0.226 ± 0.001	0.813 ± 0.001	0.075 ± 0.000	0.426 ± 0.001	0.071 ± 0.000
MIL top-2 aggregation	30 s	0.760 ± 0.001	0.218 ± 0.001	0.763 ± 0.003	0.076 ± 0.001	0.453 ± 0.009	0.078 ± 0.001
Standard multi-view	30 s	0.765 ± 0.000	0.226 ± 0.001	0.784 ± 0.002	0.084 ± 0.001	0.458 ± 0.003	0.076 ± 0.001
EMA mean-teacher	30 s	0.765 ± 0.001	0.227 ± 0.002	0.787 ± 0.008	0.084 ± 0.003	0.450 ± 0.011	0.076 ± 0.004
RWS	**30 s**	**0.769 ± 0.001**	**0.232 ± 0.001**	**0.886 ± 0.001**	**0.033 ± 0.001**	**0.331 ± 0.004**	**0.057 ± 0.001**
Self-relation	60 s	0.770 ± 0.001	0.234 ± 0.001	0.823 ± 0.009	0.074 ± 0.003	0.397 ± 0.009	0.070 ± 0.002
Teacher–student	60 s	0.774 ± 0.000	0.239 ± 0.000	0.847 ± 0.000	0.067 ± 0.000	0.379 ± 0.007	0.067 ± 0.001
MIL top-2 aggregation	60 s	0.767 ± 0.001	0.229 ± 0.002	0.794 ± 0.010	0.070 ± 0.003	0.412 ± 0.005	0.073 ± 0.001
Standard multi-view	60 s	0.771 ± 0.001	0.237 ± 0.003	0.818 ± 0.005	0.076 ± 0.003	0.400 ± 0.007	0.071 ± 0.002
EMA mean-teacher	60 s	0.770 ± 0.002	0.236 ± 0.002	0.832 ± 0.004	0.073 ± 0.004	0.389 ± 0.010	0.067 ± 0.008
RWS	**60 s**	**0.779 ± 0.001**	**0.249 ± 0.001**	**0.913 ± 0.002**	**0.032 ± 0.001**	**0.278 ± 0.007**	**0.052 ± 0.001**

Rows report absolute three-seed means ± SD for matched repeatability methods. RWS is included as its own method row; paired uncertainty is reported separately in Supplementary Table S8c.

4.2 Comparison with alternative repeatability strategies

Table 2 compares RWS with five repeatability strategies under matched BiGRU and eight-window exposure: self-relation, teacher–student, MIL top-2, standard multi-view, and EMA mean-teacher. RWS has the strongest repeatability profile across the three input durations. Longer inputs improve discrimination and repeatability for the compared methods.

The paired bootstrap shows favorable intervals for the four repeatability endpoints against each alternative across durations(Table S8c). After BH–FDR and Bonferroni correction, the repeatability comparisons remain significant(Table S9b). Calibrated Brier, ECE, slope, and intercept for this comparison are given in Table S5.

4.3 Fair exposure

Table 3 separates the effect of seeing eight windows from the effect of the RWS objective. BCE-only multi-window training does not reproduce RWS: it generally retains the Conventional repeatability profile or worsens it, whereas RWS improves ICC and lowers within-episode SD, flip rate, and VDFR. These differences persist as input duration increases. All six primary endpoints have favorable paired-bootstrap intervals at every duration against Conventional training; against BCE-only multi-window training, the repeatability gains remain clear. The corresponding interval summaries are in Table S8b. The calibrated fair-exposure comparison is reported in Table S6.

Table 3: Fair-exposure controls.

Training condition	Duration	AUROC	AUPRC	ICC	Within-episode SD	Flip rate	VDFR
Conventional	10 s	0.732 ± 0.002	0.183 ± 0.002	0.696 ± 0.001	0.094 ± 0.000	0.595 ± 0.006	0.080 ± 0.000
BCE-only multi-window	10 s	0.744 ± 0.000	0.198 ± 0.000	0.695 ± 0.006	0.100 ± 0.001	0.585 ± 0.004	0.083 ± 0.002
RWS	**10 s**	**0.740 ± 0.000**	**0.198 ± 0.001**	**0.811 ± 0.002**	**0.021 ± 0.001**	**0.458 ± 0.009**	**0.065 ± 0.001**
Conventional	30 s	0.760 ± 0.001	0.219 ± 0.002	0.797 ± 0.002	0.079 ± 0.001	0.457 ± 0.007	0.072 ± 0.000
BCE-only multi-window	30 s	0.764 ± 0.000	0.225 ± 0.001	0.784 ± 0.009	0.084 ± 0.003	0.460 ± 0.004	0.077 ± 0.002
RWS	**30 s**	**0.769 ± 0.001**	**0.232 ± 0.001**	**0.886 ± 0.001**	**0.033 ± 0.001**	**0.331 ± 0.004**	**0.057 ± 0.001**
Conventional	60 s	0.769 ± 0.001	0.232 ± 0.002	0.836 ± 0.001	0.070 ± 0.000	0.396 ± 0.008	0.067 ± 0.000
BCE-only multi-window	60 s	0.772 ± 0.000	0.238 ± 0.001	0.822 ± 0.007	0.077 ± 0.002	0.400 ± 0.011	0.071 ± 0.001
RWS	**60 s**	**0.779 ± 0.001**	**0.249 ± 0.001**	**0.913 ± 0.002**	**0.032 ± 0.001**	**0.278 ± 0.007**	**0.052 ± 0.001**

4.4 Loss ablation

The component-level test is reported in Supplementary Table S15. Removing the leave-one-out relational term, the logit-range hinge, or the probability-ICC surrogate weakens the repeatability profile relative to complete RWS at all three durations. The paired bootstrap supports the repeatability endpoint differences for each component removal. Calibrated ablation results are reported in Table S7.

4.5 Training-window and deployment-window analyses

Figure S1 shows the $K = 1$–10 training-window scan. The first clear repeatability gain appears when moving from $K = 1$ to $K = 2$; the repeatability curves change much less after $K = 8$ than during the early part of the scan. Discrimination follows the same broad pattern. We therefore use $K = 8$ as the fixed protocol. Figure S2 shows that wall-clock time and peak allocated/reserved memory increase with K; window throughput improves initially but has diminishing returns at larger banks.

Figure S3 reports the fixed-checkpoint deployment-window audit defined in the experimental protocol. From $D = 1$ to $D = 4$, AUROC and AUPRC increase for both conditions, while ICC increases and within-episode SD, flip rate, and VDFR decrease.

4.6 Cross-dataset transfer and independent replication

Frozen MC-MED checkpoints transferred to MIMIC-III-Ext-PPG with RWS improving all six primary endpoints at 10, 30, and 60 s (Table 4, Panel A). At 60 s, AUROC/AUPRC rose from 0.647/0.297 to 0.668/0.314 and ICC from 0.868 to 0.914, while within-episode SD, flip rate, and VDFR fell from 0.048/0.292/0.056 to 0.024/0.244/0.052. The direction of the calibration-dependent endpoints also held under source-frozen calibration. All 18 migration contrasts had positive subject-bootstrap intervals and survived Holm and BH–FDR correction (Supplementary Tables S13–S14).

Independent in-domain MIMIC-III-Ext-PPG retraining reproduced the ICC, within-episode SD, flip-rate, and VDFR improvements at every duration and directionally maintained or improved discrimination (Table 4, Panel B). Only the 10-s AUROC contrast had a non-significant interval, leaving 17 of 18 retraining contrasts significant after correction. The target-domain K scan showed an early gain by $K = 2$ and smaller later changes, while RWS retained its advantage at matched $D = 1, 2, 4$.

5 Discussion

Discrimination metrics do not capture whether a model produces repeatable outputs for repeated measurements from the same short-term episode. Our results demonstrate that prediction repeatability can be learned by exploiting naturally repeated windows as relational supervision. RWS leverages same-episode relationships during training while preserving per-window supervision and single-window deployment.

The fair-exposure, alternative-strategy, and ablation analyses indicate that the improvement is not explained by additional window exposure alone. By constraining within-episode prediction variation while preserving task supervision and between-episode separation, RWS improves repeatability without requiring inference-time aggregation or sacrificing discriminative ability. This design

Table 4: Cross-dataset validation on MIMIC-III-Ext-PPG.

Panel A: Frozen checkpoint transfer

Duration	Method	AUROC	AUPRC	ICC	Within-episode SD	Flip rate	VDFR
10 s	Conventional	0.616 ± 0.003	0.272 ± 0.002	0.758 ± 0.007	0.059 ± 0.001	0.462 ± 0.010	0.070 ± 0.001
10 s	**RWS**	**0.628 ± 0.002**	**0.280 ± 0.002**	**0.831 ± 0.005**	**0.017 ± 0.000**	**0.374 ± 0.005**	**0.064 ± 0.001**
30 s	Conventional	0.633 ± 0.001	0.287 ± 0.001	0.827 ± 0.001	0.053 ± 0.000	0.353 ± 0.004	0.062 ± 0.000
30 s	**RWS**	**0.652 ± 0.002**	**0.302 ± 0.001**	**0.884 ± 0.000**	**0.025 ± 0.001**	**0.294 ± 0.006**	**0.057 ± 0.001**
60 s	Conventional	0.647 ± 0.005	0.297 ± 0.003	0.868 ± 0.001	0.048 ± 0.001	0.292 ± 0.009	0.056 ± 0.002
60 s	**RWS**	**0.668 ± 0.002**	**0.314 ± 0.002**	**0.914 ± 0.002**	**0.024 ± 0.001**	**0.244 ± 0.006**	**0.052 ± 0.001**

Panel B: Independent in-domain retraining

Duration	Method	AUROC	AUPRC	ICC	Within-episode SD	Flip rate	VDFR
10 s	Conventional	0.655 ± 0.002	0.304 ± 0.003	0.761 ± 0.007	0.057 ± 0.002	0.450 ± 0.009	0.089 ± 0.002
10 s	**RWS**	**0.656 ± 0.007**	**0.311 ± 0.005**	**0.886 ± 0.004**	**0.013 ± 0.001**	**0.278 ± 0.013**	**0.063 ± 0.000**
30 s	Conventional	0.652 ± 0.004	0.302 ± 0.004	0.720 ± 0.009	0.062 ± 0.006	0.446 ± 0.009	0.084 ± 0.003
30 s	**RWS**	**0.661 ± 0.011**	**0.318 ± 0.011**	**0.905 ± 0.005**	**0.012 ± 0.002**	**0.242 ± 0.008**	**0.052 ± 0.003**
60 s	Conventional	0.652 ± 0.007	0.305 ± 0.007	0.715 ± 0.012	0.062 ± 0.003	0.436 ± 0.006	0.085 ± 0.002
60 s	**RWS**	**0.676 ± 0.004**	**0.322 ± 0.002**	**0.926 ± 0.002**	**0.016 ± 0.001**	**0.187 ± 0.020**	**0.049 ± 0.002**

Panel A reports frozen-checkpoint transfer from MC-MED; Panel B reports independent in-domain retraining on MIMIC-III-Ext-PPG. Entries are three-seed mean ± SD.

reflects a selective invariance principle: observations expected to share the same underlying state are encouraged to produce repeatable predictions, while differences between monitoring episodes remain distinguishable. The repeatability pattern across architectures, representations, and input durations suggests that the learned repeatability property is not restricted to a specific backbone or representation strategy.

The MIMIC-III-Ext-PPG analyses provide complementary evidence for the transportability of this repeatability property. Frozen MC-MED checkpoints retained the repeatability advantage of RWS after transfer to MIMIC-III-Ext-PPG, despite changes in absolute discrimination performance. Independent training within MIMIC-III-Ext-PPG reproduced the same repeatability pattern, further supporting that the observed improvement is not specific to the source dataset. Together, these findings show that the repeatability advantage of RWS persisted despite a shift in absolute discrimination performance across datasets.

The scope of this claim is limited to short-term prediction of chronic disease-background states, ICD-derived disease-background labels represent coded clinical states. Therefore, improved repeatability should not be interpreted as enforcing invariance to genuine acute physiological changes. Instead, monitoring systems should balance repeatability under unchanged states with sensitivity to clinically meaningful transitions. Future work should evaluate this framework prospectively across devices, populations, and clinical settings, while incorporating signal quality assessment and scenarios involving acute physiological changes.

6 CONCLUSION

Continuous physiological monitoring requires predictive models that are not only discriminative but also repeatable under repeated observations of the same individual. In this work, we introduce RWS, a training framework that uses same-episode windows as relational context while retaining single-window inference.

Across diverse PPG modeling settings and datasets, RWS improves prediction repeatability while maintaining or improving discrimination performance. These findings demonstrate that repeatability can be learned as a property of the prediction function rather than achieved only through inference-time aggregation.

Prediction repeatability therefore provides a complementary evaluation dimension for AI systems operating on continuously acquired physiological signals and may serve as an important consideration for developing trustworthy predictive models in repeated-observation settings.

REPRODUCIBILITY STATEMENT

The anonymous code repository for this submission is available at `https://anonymous.4open.science/r/RWSFRSWPDBP-D1F5`. It provides the active experiment matrix, runner configurations, checkpoint-analysis drivers, numeric source data, table-generation scripts, and checksums. The protocol fixes the dataset, patient-disjoint chronological split, three seeds (42, 43, 44), input durations, training window count, evaluation window count, and primary single-window deployment setting before summarization. The frozen-encoder checkpoint filenames, SHA256 identifiers, preprocessing contracts, and public model sources are recorded in the artifact provenance file. All table values in this manuscript are generated from the frozen CSV sources rather than manually transcribed; private filesystem paths are not part of the submitted manuscript.

AI USE STATEMENT

In this work, generative AI tools assisted with efforts to aid or polish writing. The reported model-training runs were completed before manuscript assembly; the tools did not alter model checkpoints or training measurements. The authors reviewed and tested AI-assisted text and remain responsible for the scientific claims, numerical sources, citations, anonymity, and final submission. The compiled PDF has been checked, and the same disclosure is provided in the submission form.

REFERENCES

Zachi I. Attia, Suraj Kapa, Francisco Lopez-Jimenez, Paul M. McKie, David J. Ladewig, Gauri Satam, Patricia A. Pellikka, Maurice Enriquez-Sarano, Peter A. Noseworthy, Thomas M. Munger, Samuel J. Asirvatham, Christopher G. Scott, Rick E. Carter, Xinyan Yao, and Paul A. Friedman. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. *Nature Medicine*, 25:70–74, 2019a. doi: 10.1038/s41591-018-0240-2.

Zachi I. Attia, Peter A. Noseworthy, Francisco Lopez-Jimenez, Samuel J. Asirvatham, Anil J. Deshmukh, Bernard J. Gersh, David J. Ladewig, Gauri Satam, Paul A. Friedman, et al. An artificial intelligence-enabled ecg algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *The Lancet*, 394(10201):861–867, 2019b. doi: 10.1016/S0140-6736(19)31721-0.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. Moment: A family of open time-series foundation models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 16115–16152, 2024. URL `https://arxiv.org/abs/2402.03885`.

Xiao Gu, Wei Tang, Jinpei Han, Veer Sangha, Fenglin Liu, Shreyank N. Gowda, Antonio H. Ribeiro, Patrick Schwab, Kim Branson, Lei Clifton, et al. Cardiac health assessment across scenarios and devices using a multimodal foundation model pretrained on data from 1.7 million individuals. *Nature Machine Intelligence*, 8(2):220–233, 2026. doi: 10.1038/s42256-026-01180-5.

Y. Guo, H. Wang, H. Zhang, et al. Photoplethysmography-based machine learning approaches for atrial fibrillation prediction: A report from the huawei heart study. *JACC: Asia*, 1(3):399–408, 2021. doi: 10.1016/j.jacasi.2021.09.004.

Fredrik K. Gustafsson, Xiao Gu, Mattia Carletti, Patitapaban Palo, David W. Eyre, and David A. Clifton. Signalmc-med: A multimodal benchmark for evaluating biosignal foundation models on single-lead ecg and ppg. *arXiv preprint arXiv:2603.09940*, 2026. URL `https://arxiv.org/abs/2603.09940`.

S. Hamada, M. Amemiya, M. Ochida, et al. Temporal variability and influence of measurement conditions of ai-based atrial fibrillation risk estimation. *Journal of Arrhythmia*, 42(1):e70280, 2026. doi: 10.1002/joa3.70280.

Awni Y. Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, 25:65–69, 2019. doi: 10.1038/s41591-018-0268-3.

Aman Kansal, Emma Chen, Boyang Tom Jin, Pranav Rajpurkar, and David A. Kim. Mc-med, multimodal clinical monitoring in the emergency department. *Scientific Data*, 12:1094, 2025. doi: 10.1038/s41597-025-05419-5.

Jun Li, Aaron D. Aguirre, Valdery Moura, Jiarui Jin, Che Liu, Lanhai Zhong, Chenxi Sun, Gari Clifford, M. Brandon Westover, and Shenda Hong. An electrocardiogram foundation model built on over 10 million recordings. *NEJM AI*, 2(7):AIoa2401033, 2025. doi: 10.1056/AIoa2401033. URL https://ai.nejm.org/doi/10.1056/AIoa2401033.

S. C. McLaughlin, T. C. Aitchison, T. F. Yang, and P. W. Macfarlane. Smoothing techniques for improving the repeatability of the automated diagnosis of inferior myocardial infarction. *Journal of Electrocardiology*, 27(Suppl):182–187, 1994. doi: 10.1016/S0022-0736(94)80089-8.

S. C. McLaughlin, P. Chishti, T. C. Aitchison, and P. W. Macfarlane. Techniques for improving overall consistency of serial ecg analysis. *Journal of Electrocardiology*, 29(Suppl):41–45, 1996. doi: 10.1016/S0022-0736(96)80018-5.

Mohammad Moulaeifard, Marie Kutscher, Philip J. Aston, Peter H. Charlton, and Nils Strodthoff. Mimic-iii-ext-ppg, a ppg-based benchmark dataset for cardiovascular and respiratory signal analysis. *Scientific Data*, 13:668, 2026. doi: 10.1038/s41597-026-07335-8.

Arvind Pillai, Dimitris Spathis, Fahim Kawsar, and Mohammad Malekzadeh. Papagei: Open foundation models for optical physiological signals. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.20542.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Geoffrey H. Tison, Jonathan M. Sanchez, Brandon Ballinger, et al. Passive detection of atrial fibrillation using a commercially available smartwatch. *JAMA Cardiology*, 3:409–416, 2018. doi: 10.1001/jamacardio.2018.0136.