RESEARCH ARTICLE

BULLETIN OF SCIENTIFIC RESEARCH

Asian Research Association

DOI: 10.54392/bsr2511

Motion-Robust Heart-Rate Estimation from Wrist Photoplethysmography and Tri-Axial Acceleration: A Participant-Independent Evaluation on the PPG-DaLiA Dataset

R. Kishore Kanna [1, *]

[1] Department of Biomedical Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Tamil Nadu, India

*Corresponding author Email: kishorekanna007@gmail.com

DOI: https://doi.org/10.54392/bsr2613

Received: 19-01-2026; Revised: 12-03-2026; Accepted: 25-03-2026; Published: 10-04-2026

Abstract: Continuous heart-rate (HR) monitoring from wrist-worn photoplethysmography (PPG) is central to consumer and clinical wearable devices, yet motion artefact remains the principal obstacle to reliable estimation during daily activity. This study develops and rigorously evaluates a motion-aware, feature-based machine-learning pipeline for HR estimation from synchronised wrist PPG and tri-axial accelerometry, using the publicly available PPG-DaLiA dataset (15 participants, Empatica E4 wrist sensor, ECG-derived reference HR). Eight-second, two-second-shifted windows were constructed from the deposited signals (64,697 windows retained), and eighteen spectral and time-domain descriptors were extracted from the PPG and acceleration-magnitude spectra. Three regression models (ridge regression, Extra-Trees, histogram-based gradient boosting) were compared under five-fold, participant-grouped cross-validation restricted to a development partition of ten participants, the Extra-Trees regressor was selected by out-of-fold participant-macro mean absolute error (MAE) and frozen before any exposure to the five held-out test participants. On the untouched test partition, the frozen model achieved an MAE of 14.12 bpm (95% participant-clustered bootstrap CI, reported in-text) against a deterministic spectral-peak baseline's MAE of 24.61 bpm, a mean improvement of 10.49 bpm (95% CI −18.41 to −4.02), with concordance correlation coefficient (CCC) of 0.567 versus 0.193 for the baseline. Permutation importance identified the dominant PPG spectral peak frequency, secondary PPG peak, and PPG spectral entropy as the leading predictors. An exploratory one-dimensional convolutional neural network (CNN), trained only for comparison and never permitted to replace the frozen classical model, achieved a lower test MAE of 10.93 bpm, suggesting head-room for representation-learning approaches while confirming that the engineered-feature pipeline provides a robust, computationally inexpensive, and fully auditable alternative. All modelling decisions were made without access to the test partition, and a leakage audit confirmed no participant overlap between partitions. These results quantify, with participant-independent rigour and calibrated uncertainty, the accuracy achievable by classical ensemble learning on engineered spectral features for wrist-PPG heart-rate estimation, and they identify motion artefact — rather than model capacity — as the primary remaining source of error.

Keywords: Photoplethysmography, Heart-Rate Estimation, Wearable Sensors, Motion Artefact, Extra-Trees, PPG-Dalia, Feature Engineering, Participant-Independent Validation.

1. Introduction

Heart rate (HR) is among the most widely monitored physiological signals in consumer health technology, underpinning applications that range from fitness tracking and sleep staging to early detection of arrhythmia and cardiovascular risk stratification. Photoplethysmography (PPG)—an optical technique that infers blood-volume changes from light absorbed or reflected by cutaneous vasculature—has become the dominant sensing modality for continuous, wrist-worn HR monitoring because it is inexpensive, low-power, and unobtrusive relative to electrocardiography (ECG) [1]. Nearly every commercial smartwatch and fitness band, from research-grade devices such as the Empatica E4 to mass-market products, now relies on PPG for continuous cardiac monitoring.

Despite its ubiquity, PPG-based HR estimation remains fundamentally more difficult than ECG-based estimation because the optical signal is highly susceptible to motion artefact: mechanical displacement of the sensor relative to the skin, changes in local blood flow induced by muscular contraction,

and ambient light leakage can corrupt the waveform and obscure the cardiac-synchronous pulsatile component. In a controlled comparison across six wearable devices, no statistically significant overall difference in HR accuracy was observed across skin tones, however, device type and activity condition significantly affected error, and absolute error during activity was, on average, 30% higher than during rest [2]. During walking, cycling, resistance exercise, and other periodic activities, motion-related spectral components can overlap or dominate the cardiac component, making naive spectral-peak selection unreliable [3]. This motion sensitivity is not a peripheral engineering nuisance, it is the central open problem in wearable cardiac sensing and directly limits the clinical and consumer trustworthiness of PPG-derived HR.

The consequences of this fragility are not merely academic. Wearable-derived HR streams increasingly feed into consumer-facing risk scores, fitness recommendations, and, in an expanding number of regulatory-cleared products, screening algorithms for atrial fibrillation and other arrhythmias, systematic error during precisely the activity periods when a user is most engaged with the device — exercising, commuting, performing manual work — undermines the reliability of every downstream computation built on top of the raw HR estimate. Addressing this fragility requires either explicitly modelling and subtracting the motion-induced interference at the signal-processing level, or training a data-driven estimator that implicitly learns to discount motion-corrupted spectral content. This study pursues the second strategy while remaining conceptually and empirically grounded in the accelerometer-referenced insight that underlies the first.

A substantial body of signal-processing research has addressed motion-artefact compensation by exploiting a simultaneously recorded accelerometer as a reference channel for the interfering motion. The TROIKA framework formalised this approach through signal decomposition, sparse spectral reconstruction, and spectral peak tracking with verification. A subsequent joint sparse spectrum reconstruction method used accelerometer-referenced spectral subtraction to improve computational efficiency while retaining strong accuracy [4]. These classical pipelines were frequently tuned to the small IEEE Signal Processing Cup datasets on which they were originally validated, later work on the larger and more activity-diverse PPG-DaLiA benchmark therefore examined robustness and generalisation under conditions closer to daily life [5].

PPG-DaLiA comprises synchronised wrist PPG, wrist tri-axial acceleration, and chest ECG-derived reference HR from fifteen participants performing daily-life activities including sitting, walking, stair ascent and descent, driving, table work, cycling, and light resistance exercise [6]. The recordings were acquired with consumer-representative hardware (Empatica E4 at the wrist and RespiBAN at the chest), and the dataset has become a widely used benchmark for evaluating HR estimation under heterogeneous motion conditions. The original Deep PPG study evaluated convolutional neural networks on time-frequency representations, while a subsequent multi-headed convolutional-LSTM architecture reported a participant-averaged MAE of 6.28 bpm and a Pearson correlation of 0.85 on PPG-DaLiA [7]. Together, these studies demonstrate both the value of learned representations and the marked participant-to-participant variability of this benchmark.

Two methodological gaps motivate the present study. First, much of the literature evaluating machine-learning approaches to PPG-based HR estimation reports performance without transparently documenting protections against information leakage across the boundary between model development and final evaluation — a boundary that is only meaningful if enforced at the participant level, since windows from the same individual share physiology, sensor placement, and activity context that a model can exploit to appear more accurate than it would be for a genuinely new wearer. Second, feature-based classical ensemble methods — despite being computationally inexpensive, fully interpretable, and readily deployable on resource-constrained wearable hardware — are comparatively under-represented in recent PPG-DaLiA benchmarking relative to deep architectures, even though their accuracy–complexity trade-off is directly relevant to on-device inference.

This study addresses both gaps through a single, prespecified analytical pipeline applied to the official PPG-DaLiA release. We (i) construct a motion-aware spectral and time-domain feature set from synchronised wrist PPG and accelerometry, (ii) select among three regression architectures—ridge regression, Extra-Trees [8], and histogram-based gradient boosting [9]—using exclusively development-partition, participant-grouped cross-validation, (iii) freeze the selected model before any exposure to a held-out test partition of five participants, (iv) quantify held-out accuracy against a deterministic spectral-peak baseline using participant-clustered bootstrap confidence intervals, Bland-Altman agreement analysis [10], and the concordance correlation coefficient [11], (v) interpret the frozen model through permutation feature importance as introduced for random forests [12] and subsequently formalised as model reliance [13], and (vi) benchmark the classical pipeline against an exploratory one-dimensional convolutional neural network that is explicitly barred from replacing the frozen primary model regardless of its comparative performance. Throughout, every modelling and selection decision is documented with an explicit leakage audit, so that the reported test-set accuracy

reflects generalisation to genuinely unseen participants rather than to unseen windows drawn from already-observed individuals. The objective is not merely to report an accuracy figure, but to provide a methodologically transparent, participant-independent, and statistically calibrated account of what a lightweight, auditable machine-learning pipeline can and cannot achieve for motion-robust wrist-PPG heart-rate estimation, and to identify—through interpretation and comparison with an exploratory deep model—where the remaining error originates.

Beyond its technical contribution, this question has practical significance for the growing ecosystem of consumer and clinical-adjacent wearables that rely on PPG-derived HR as an input to downstream algorithms — activity classification, sleep staging, stress and recovery scoring, and, increasingly, screening for irregular rhythms. Errors in the underlying HR estimate propagate into every one of these downstream computations, and because motion artefact correlates with exactly the physical activity that many wearable applications are designed to monitor, the periods of greatest algorithmic interest are frequently also the periods of greatest measurement unreliability. A transparent, participant-independent account of achievable accuracy — together with a clear description of which signal properties a model relies upon and where its errors concentrate — is therefore of direct value not only to the signal-processing and machine-learning research community, but to developers and evaluators of downstream wearable-health applications who must decide how much confidence a given HR stream can support.

The remainder of this paper is organised as follows. Section 2 reviews the relevant literature on PPG motion-artefact compensation, machine-learning approaches evaluated on PPG-DaLiA, ensemble tree methods and their interpretation, and statistical practice for agreement studies in wearable-sensor validation. Section 3 details the dataset, preprocessing, feature engineering, participant-level partitioning, model development and selection, statistical evaluation framework, model interpretation, exploratory deep-learning comparator, and reproducibility controls. Section 4 reports dataset characteristics, development-only model selection, held-out test-partition evaluation, agreement analysis, feature importance, and the exploratory comparator's results. Section 5 discusses the principal findings, their mechanistic interpretation, their relationship to prior PPG-DaLiA benchmarks, the trade-off between classical and deep approaches, and the broader methodological implications of participant-independent evaluation. Section 6 states the limitations of the study, and Section 7 concludes.

2. Related Work

2.1 Photoplethysmography and the challenge of motion artefact

Photoplethysmography measures the pulsatile modulation of transmitted or reflected light caused by cyclical variation in cutaneous blood volume, and under stationary conditions the dominant frequency component of the waveform corresponds closely to the cardiac cycle. At the wrist, however, the light path traverses a thin, mechanically mobile region of tissue, so the signal-to-noise ratio is markedly worse than at better-perfused sites such as the fingertip or earlobe. The wearable-photoplethysmography roadmap identifies motion-related signal-quality degradation as a principal unresolved barrier to extending PPG beyond simple heart-rate tracking. Controlled device comparisons likewise show higher error during activity and significant variation between devices, while not supporting a statistically significant overall difference in HR accuracy across skin tones. These findings indicate that motion artefact and signal quality, rather than sensor calibration alone, remain major sources of inaccuracy in wearable optical HR.

Two broad families of technique have been proposed to compensate for motion artefact. The first is adaptive and decomposition-based signal processing that exploits a synchronously recorded accelerometer as a reference for the corrupting motion, the second, more recent family applies supervised machine learning — including deep neural networks — directly to raw or lightly processed multimodal sequences.

2.2 Accelerometer-Referenced Signal-Processing Approaches

The TROIKA framework is the most widely cited representative of the first family. It decomposes the PPG signal to suppress motion-correlated components, reconstructs a high-resolution spectrum via sparse signal recovery, and then tracks the spectral peak corresponding to HR across successive windows with a verification step designed to reject implausible frequency jumps. Zhang subsequently proposed a joint sparse spectrum reconstruction (JOSS) variant that performs spectral subtraction using the accelerometer channels rather than full signal decomposition, improving computational efficiency while retaining strong accuracy on the same benchmark. These and related methods were predominantly validated on the IEEE Signal Processing Cup datasets, which comprise recordings from only twelve to twenty-three subjects during controlled running protocols, because the datasets are small and the activity protocol is comparatively homogeneous (treadmill running at prescribed speeds), the reported per-subject MAE figures — often below 2–3 bpm — may not transfer

directly to the more heterogeneous, closer-to-real-life activity mix captured in larger datasets such as PPG-DaLiA.

The present study adopts the conceptual core of the accelerometer-referenced tradition — using the accelerometer spectrum to characterise and, implicitly, discount motion energy that overlaps the PPG spectrum — but implements it through supervised feature-based regression rather than an explicit decomposition-and-tracking algorithm. This choice trades some of the elegance of a fully deterministic tracking framework for the flexibility of a learned, nonlinear mapping between spectral descriptors of both channels and the reference HR, and for straightforward comparability against a deterministic spectral-peak baseline evaluated on the same partitioned data.

2.3 Machine-Learning and Deep-Learning Approaches On PPG-Dalia

Reiss et al. introduced PPG-DaLiA alongside a convolutional neural network operating on time-frequency spectrograms of the synchronised PPG and accelerometer channels, reporting that the deep-learning approach reduced MAE by approximately 31% relative to their extended classical baseline on PPG-DaLiA. This result established the dataset as a benchmark for demonstrating the generalisation advantages of learned representations over fixed spectral-tracking heuristics when sufficient training data are available. Wilkosz and Szczęsna extended this line of work with a multi-headed convolutional–LSTM architecture (MH Conv-LSTM DeepPPG) that processes PPG and accelerometer channels through separate convolutional heads before recurrent temporal integration, reporting a participant-averaged MAE of 6.28 bpm and a Pearson correlation of 0.85 against the ECG-derived reference on PPG-DaLiA, the authors additionally reported that error was highly participant-dependent, with some participants (notably one with concurrent tachycardia) showing markedly worse agreement than others — an observation that anticipates the participant-level heterogeneity we also document in Section 4.

Relative to this literature, the present study makes three deliberate choices. First, it restricts model and hyperparameter selection entirely to a development partition, evaluating the frozen model exactly once on an untouched test partition — a discipline that is stated in several prior studies but not always independently auditable from the reported methodology. Second, it emphasises a classical, feature-based ensemble (Extra-Trees) as the primary model, prioritising interpretability, computational economy, and auditability over the marginal accuracy gains that deep architectures can offer, while still including an exploratory CNN comparator to characterise the size of that gap under our own partitioning and preprocessing. Third, it reports calibrated uncertainty — participant-clustered bootstrap confidence intervals for all headline metrics — rather than point estimates alone, which is comparatively rare in this literature despite the substantial participant-to-participant variability documented both here and by Wilkosz and Szczęsna.

2.4 Ensemble Tree Methods and Their Interpretation

The Extra-Trees regressor extends the random-forest principle by randomising both the splitting attribute and the cut-point at each node, which typically reduces variance relative to standard bagged trees at modest computational cost and has proven competitive across a wide range of tabular regression tasks. Histogram-based gradient boosting instead builds an additive ensemble of shallow trees trained sequentially to correct the residual error of the ensemble so far, it is frequently highly accurate on structured, engineered-feature data and was included here as a second, methodologically distinct tree-ensemble candidate. Ridge regression served as a linear reference model to establish how much of the achievable accuracy on this feature set is attributable to nonlinear interaction effects that only the tree ensembles can capture. To interpret the frozen ensemble, we use permutation feature importance, which quantifies the increase in prediction error when a single feature's values are randomly shuffled, thereby breaking its association with the outcome while leaving the joint distribution of the remaining features intact. Fisher, Rudin, and Dominici formalised and generalised this idea as "model reliance," clarifying its interpretation as an estimate of how much a specific fitted model depends on a given covariate for its accuracy, and cautioning that importance rankings are model-specific rather than universal properties of the underlying data-generating process — a caveat we revisit when interpreting our results in Section 5.

2.5 Statistical Evaluation of Agreement in Physiological Measurement

Reporting a single point estimate of error is insufficient for a measurement-agreement study, particularly when only a small number of held-out participants contribute a large but statistically dependent set of windows. The Bland-Altman method remains the standard approach in clinical measurement science for visualising and quantifying agreement between a candidate measurement and a reference through the mean difference (bias) and the 95% limits of agreement. The concordance correlation coefficient complements Pearson correlation by simultaneously penalising both imprecision and

systematic location or scale shift relative to the identity line. Because repeated windows from the same participant are not statistically independent, we apply the general bootstrap principle introduced by Efron [14] at the participant level: participants, rather than individual windows, are resampled with replacement when constructing confidence intervals and paired performance contrasts.

2.6 Synthesis and Positioning of the Present Study

Taken together, the literature reviewed above establishes three points that directly inform the design of this study. First, accelerometer-referenced compensation is the dominant strategy for addressing PPG motion artefact, whether implemented through explicit signal decomposition or through learned representations that implicitly weight accelerometer information, this study occupies a deliberate middle ground by operationalizing the same principle through interpretable engineered features. Second, dataset scale and activity diversity materially affect how well reported accuracy figures are likely to generalise, and PPG-DaLiA is a more heterogeneous benchmark than the small, single-activity datasets used in much of the classical signal-processing literature. Third, PPG-DaLiA studies do not uniformly report participant-level leakage controls and clustered uncertainty quantification. The present study is designed specifically to address this methodological gap while remaining directly comparable, through a matched exploratory deep-learning comparator, to accuracy levels reported elsewhere.

3. Materials and Methods

Figure 1 summarises the complete analytical pipeline described in this section, from raw signal acquisition through participant-level partitioning, development-only model selection, frozen held-out evaluation, interpretation, and the exploratory deep-learning comparator.

3.1 Dataset

This study uses the PPG-DaLiA dataset in its entirety, obtained directly from the official UCI Machine Learning Repository release (UCI dataset ID 495, DOI 10.24432/C53890), distributed under a CC BY 4.0 licence. PPG-DaLiA contains recordings from fifteen participants (S1–S15), each performing a semi-structured protocol of daily-life activities for approximately 150 minutes, including sitting, ascending/descending stairs, table work, walking, driving a car, cycling, working, and light resistance exercises, interspersed with brief transitions. Two physiological signal streams are used in this study: (i) wrist blood-volume-pulse (BVP), a PPG-derived signal recorded at 64 Hz by the wrist-worn Empatica E4 device, and (ii) wrist tri-axial acceleration (ACC), recorded at 32 Hz by the same device. The dataset additionally provides ECG-derived reference heart-rate labels, computed from a chest-worn RespiBAN device and deposited alongside the raw signals at a fixed labelling cadence, these reference labels — not the raw ECG waveform itself — served as the regression target throughout, consistent with the intended use of the deposited data. No ECG signal was used as a predictor at any stage. Prior to analysis, the official archive was downloaded programmatically, its SHA-256 digest was computed and recorded for provenance, and the archive was safely extracted with explicit path-traversal checks. The pipeline required exactly the fifteen original per-participant pickle files (S1–S15), the presence of all fifteen was verified before proceeding, and a per-participant file-integrity manifest (byte count and SHA-256 digest) was generated and retained as a reproducibility artefact. No participant records were excluded, added, or otherwise substituted.

3.1.1 Ethical and data-governance considerations

PPG-DaLiA is a previously collected, publicly deposited, de-identified dataset released under a CC BY 4.0 licence by its original creators, who are responsible for the human-subjects approvals and consent procedures associated with the original data collection. This study performs a secondary, fully retrospective analysis of the deposited signals and reference labels, no new human-subjects data were collected, no attempt was made to re-identify participants, and no participant-identifying metadata beyond the anonymised study codes (S1–S15) already present in the public release were used at any stage. Because the analysis is restricted to secondary use of de-identified, publicly available data under its original licence terms, no additional institutional ethical approval was sought for this specific analysis, consistent with standard practice for secondary analyses of openly licensed, de-identified physiological datasets.

3.2 Signal Preprocessing and Window Construction

Raw wrist BVP and tri-axial ACC sequences were deposited at different native sampling rates (64 Hz and 32 Hz respectively). Because the accelerometer signal was the lower-rate channel and the more limiting factor for windowed spectral analysis, the BVP channel was downsampled to 32 Hz by pairwise averaging of adjacent samples, an anti-aliasing-consistent operation that preserves the low-frequency (sub-4 Hz) spectral content relevant to the cardiac band while aligning the two modalities on a common sampling grid.

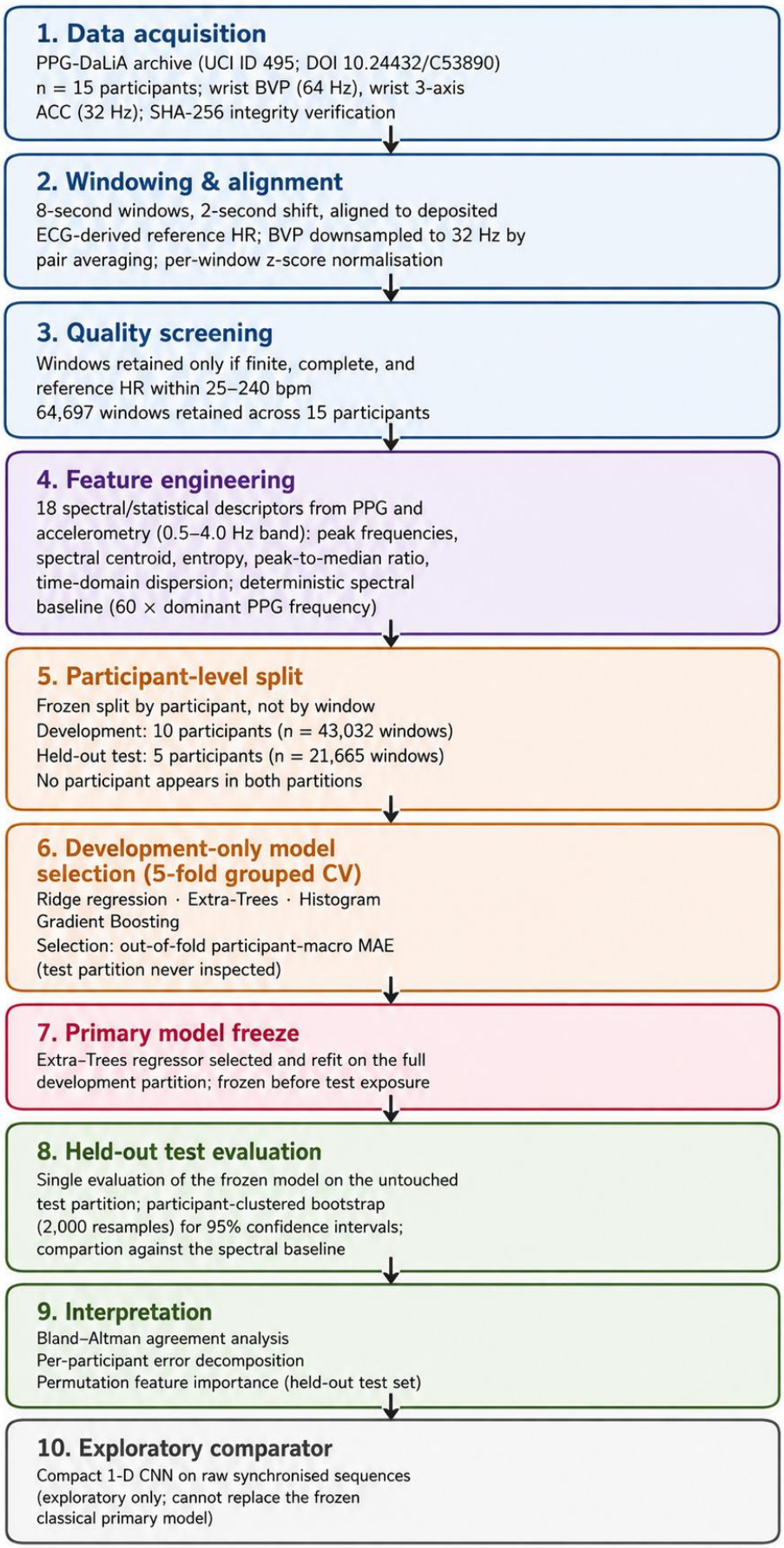

Figure 1. Analytical pipeline for motion-robust wrist-PPG heart-rate estimation, from raw signal acquisition through participant-level partitioning, development-only model selection, frozen held-out evaluation, interpretation, and the exploratory deep-learning comparator.

Analysis windows of 8 seconds duration were then constructed with a 2-second shift between consecutive windows, following the reference-label cadence deposited with the dataset, this window length and shift are standard in the PPG motion-artefact literature because they balance sufficient spectral resolution in the cardiac band against the assumption of approximately stationary HR within a window. An 8-second window at a 32 Hz analysis rate yields 256 samples per channel, giving a fast Fourier transform frequency resolution of 1/8 Hz (0.125 Hz), which is sufficient to distinguish adjacent cardiac frequencies at the resolution relevant to bpm-level HR reporting (a 0.125 Hz bin corresponds to 7.5 bpm at the transform's native resolution, refined further by the Hann-windowed peak-localisation procedure described below), shorter windows would degrade this resolution and increase estimate variance, while substantially longer windows would risk violating the approximately stationary-HR assumption during rapid activity transitions, such as the onset of stair climbing or cycling.

Because consecutive analysis windows overlap substantially (a 2-second shift against an 8-second window corresponds to 75% overlap), windows drawn from the same participant are highly autocorrelated in both their signal content and their prediction errors, this dependency structure is a further reason, beyond simple between-participant variation, why the participant-level partitioning and participant-clustered bootstrap procedures described in Sections 3.4 and 3.6 are necessary for valid statistical inference — treating overlapping, autocorrelated windows as independent observations would understate the true uncertainty of any reported metric considerably more than participant-level dependence alone. For each window, the corresponding deposited reference HR was retrieved by index alignment with the labelling cadence. A window was retained only if (i) both the PPG and three-axis accelerometer segments contained the expected number of samples for an 8-second window, (ii) all values in the resulting four-channel sequence (one PPG channel plus three acceleration axes) were finite, and (iii) the reference HR was finite and fell within a physiologically plausible range of 25–240 bpm. Each retained four-channel window was z-score normalised independently per channel (subtracting the channel mean and dividing by the channel standard deviation, computed within that window) prior to feature extraction and prior to use as raw input to the exploratory CNN described in Section 3.7. This preprocessing yielded 64,697 retained windows spanning all fifteen participants, with reference HR ranging from 41.7 to 187.0 bpm across the full dataset (mean 89.4 bpm), confirming that the retained sample captures both resting and vigorous-exercise physiology.

3.3 Feature Engineering

Eighteen spectral and time-domain descriptors were computed independently for each retained window from the normalised PPG channel and the Euclidean norm (magnitude) of the tri-axial acceleration channel. Power spectral estimates were obtained via a Hann-windowed real fast Fourier transform restricted to the 0.5–4.0 Hz band (equivalently, 30–240 bpm), which comfortably spans the physiological HR range observed in the dataset while excluding very-low-frequency drift and higher harmonics that are of limited discriminative value at this window length.

From the PPG spectrum, we computed: the three most prominent, mutually separated spectral peaks (ppg_peak1_hz, ppg_peak2_hz, ppg_peak3_hz, selected greedily with a minimum 0.10 Hz separation to avoid selecting adjacent bins of the same physical peak), the power-weighted spectral centroid (ppg_centroid_hz), the Shannon spectral entropy of the normalised power spectrum (ppg_entropy), which characterises how concentrated versus diffuse the spectral energy is, and the ratio of the maximum to the median spectral power (ppg_peak_ratio), which characterises the prominence of the dominant peak relative to the spectral noise floor. From the accelerometer-magnitude spectrum, the analogous quantities were computed (motion_peak_hz, motion_centroid_hz, motion_entropy, motion_peak_ratio), together with the absolute frequency gap between the dominant PPG peak and the dominant motion peak (ppg_motion_peak_gap_hz), which directly encodes the degree to which the two spectra are competing for the same frequency band — a key motion-artefact indicator, since a PPG peak that coincides with a strong motion peak is the classical failure mode of naïve spectral HR estimation. Finally, time-domain dispersion statistics were computed for both channels: standard deviation, interquartile range, and root-mean-square of the first difference (a proxy for high-frequency signal energy), for the PPG channel (ppg_sd, ppg_iqr, ppg_diff_rms) and, together with the channel mean, for the acceleration magnitude (motion_mean, motion_sd, motion_iqr, motion_diff_rms). Table 1 reports descriptive statistics for all eighteen features across the full retained sample. A deterministic spectral baseline HR estimate (60 × the dominant PPG spectral peak frequency in Hz) was additionally computed for every window, this baseline received no participant-specific or window-specific tuning and was used solely as a comparison point for the primary model at test time (Section 3.6), not as an input feature to any regression model.

The choice to summarise each window with a fixed, closed-form set of descriptors rather than feeding raw or lightly filtered waveforms directly to a classical

regressor reflects three considerations specific to the classical-ensemble modelling strategy adopted in Section 3.5. First, tree-ensemble methods such as Extra-Trees and gradient boosting operate on tabular, fixed-dimensionality inputs and have no intrinsic mechanism for exploiting the temporal or spectral structure of a raw sequence, meaningful signal structure must therefore be exposed to the model through explicit feature construction. Second, restricting the feature set to eighteen physiologically motivated, closed-form descriptors — rather than, for example, the full binned power spectrum — keeps the resulting model directly interpretable through permutation importance (Section 3.7), since each feature has a clear signal-processing meaning that can be related back to the underlying motion-artefact mechanism. Third, a compact feature set reduces the risk of the ensemble overfitting to high-dimensional, participant-specific spectral idiosyncrasies during development-partition training, which is of particular concern given that the development partition itself comprises only ten distinct participants.

3.4 Participant-level data partitioning

Because windows drawn from the same participant share physiology, sensor placement, skin characteristics, and activity sequencing, evaluating a model on windows from participants also present in its training data would systematically overstate its ability to generalise to a genuinely new wearer. All partitioning in this study was therefore performed at the participant level, never at the window level. The fifteen participant identifiers were split, using a fixed random seed, into a development partition of ten participants (S3, S4, S6, S7, S10, S11, S12, S13, S14, S15, 43,032 windows) and a held-out test partition of five participants (S1, S2, S5, S8, S9, 21,665 windows). The two partitions were verified programmatically to share no participant identifiers. Mean reference HR was closely matched between partitions (89.35 bpm in development versus 89.58 bpm in test), though the test partition exhibited somewhat higher HR variability (SD 25.70 bpm versus 21.23 bpm in development, Table 2), reflecting the particular activity mix and physiological range contributed by the five test participants under the fixed random assignment.

Within the development partition only, five-fold cross-validation was performed with folds constructed by grouping on participant identifier (i.e., each fold's validation set comprised entire held-back development participants, never a subset of windows from a participant otherwise present in that fold's training set). This grouped cross-validation was used exclusively for model and hyperparameter selection (Section 3.5), the test partition was not accessed, inspected, or used in any capacity until the single, final evaluation reported in Section 4.2.

3.5 Model Development and Selection

Three regression architectures were compared: (i) ridge regression with L2 penalty (alpha = 10.0), applied after median imputation and standardisation, serving as a linear reference, (ii) an Extra-Trees regressor with 450 trees, minimum leaf size of 3, and 80% feature subsampling per split, applied after median imputation, and (iii) a histogram-based gradient boosting regressor with 280 boosting iterations, learning rate 0.06, maximum 31 leaf nodes per tree, minimum leaf size 25, and L2 regularisation of 1.0, also applied after median imputation. All three pipelines were implemented with scikit-learn [15]. Hyperparameters were fixed a priori based on standard defaults and light manual tuning during preliminary, development-only exploration, no hyperparameter search used the test partition at any point.

For each candidate model, out-of-fold predictions were generated across the five development folds described in Section 3.4, and four summary statistics were computed from these out-of-fold predictions: the participant-macro MAE (the mean of per-participant MAE values, which weights each development participant equally regardless of how many windows they contributed — our primary selection criterion, chosen specifically to prevent participants with disproportionately many windows from dominating the selection signal), the pooled MAE, the pooled root-mean-square error (RMSE), and the coefficient of determination (R^2). The model selection rule was fixed in advance: the candidate with the lowest out-of-fold participant-macro MAE was designated the primary model. The selected model was then refit on the entire development partition (all ten participants, with no further validation held out) and serialised, this refit, frozen model — and no other — was subsequently evaluated on the test partition.

The decision to prioritise participant-macro MAE over pooled MAE as the selection criterion merits explicit justification, since the two quantities can, in principle, favour different models. Pooled MAE weights every window equally regardless of which participant contributed it, so a model that performs very well on participants who happen to contribute disproportionately many windows (for example, because they completed the activity protocol more slowly, or because more of their recording passed quality screening) could achieve a low pooled MAE while performing poorly for other, less-represented development participants. Participant-macro MAE instead weights every participant equally, directly reflecting the deployment-relevant question of how the model is expected to perform for a typical new wearer, rather than for a typical window drawn from the specific, imbalanced development sample at hand. We report both quantities in Table 3 for transparency, but

participant-macro MAE was the sole criterion used for the prespecified selection decision.

3.6 Held-Out Test Evaluation and Statistical Framework

The frozen primary model generated a single set of predictions on the untouched test partition (five participants, 21,665 windows), these predictions were compared against both the ECG-derived reference HR and the deterministic spectral baseline described in Section 3.3. Eight metrics were computed for each of the primary model and the baseline: MAE, RMSE, median absolute error, Pearson correlation coefficient, concordance correlation coefficient, R^2, mean absolute percentage error (MAPE), and the proportion of windows with absolute error at or below 5 bpm and at or below 10 bpm (accuracy-within-threshold metrics commonly reported in the wearable-validation literature).

To account for the statistical dependence among windows contributed by the same participant, 95% confidence intervals for every metric were constructed using a participant-clustered bootstrap: in each of 2,000 iterations, the five test participants were resampled with replacement, all windows belonging to the resampled participant set were pooled, and each metric was recomputed on the pooled resample, the 2.5th and 97.5th percentiles of the resulting bootstrap distribution defined the 95% confidence interval, using the participant-level resampling scheme described in Section 2.5. The same clustered-resampling procedure was applied jointly to the primary-model and baseline predictions within each bootstrap iteration to construct paired contrasts (primary minus baseline) for MAE, RMSE, and the within-10-bpm accuracy proportion, allowing direct, appropriately correlated inference on the difference in performance between the two approaches rather than relying on the difference between two independently constructed intervals.

Agreement between the primary model's estimates and the ECG-derived reference was additionally visualised using a Bland–Altman plot, reporting the mean bias and 95% limits of agreement (bias ± 1.96 × SD of the paired differences) computed on the full test-partition sample, together with a hexbin density scatter of estimated against reference HR.

3.7 Model interpretation

Held-out permutation feature importance was computed for the frozen primary model on the test partition, using twelve repeats per feature and mean absolute error as the scoring criterion, the reported importance for each feature is the mean increase in test-set MAE observed when that feature's values are independently permuted across test windows, holding the fitted model and all other feature values fixed. Because permutation importance is a property of a specific fitted model evaluated on a specific dataset rather than a universal statement about the underlying physiology, we interpret the resulting ranking as a description of which engineered descriptors this particular Extra-Trees model relies upon for its held-out accuracy, not as an independent claim about physiological causation.

3.8 Exploratory Deep-Learning Comparator

To characterise the accuracy achievable by a representation-learning approach under an identical partitioning and preprocessing protocol, a compact one-dimensional convolutional neural network was trained directly on the raw, z-score-normalised, four-channel (PPG plus tri-axial acceleration) window sequences, without the engineered spectral features of Section 3.3. The architecture comprised three one-dimensional convolutional blocks (24, 48, and 64 filters, kernel sizes 9, 7, and 5, with strides and max-pooling for temporal downsampling, each followed by batch normalisation), global average pooling, a 32-unit fully connected layer with dropout (rate 0.20), and a single linear output unit. The network was trained with the Adam optimiser (learning rate 1×10^{-3}) to minimise mean absolute error on z-score-normalised HR targets, using two development participants (S14, S15) as an internal validation set for early stopping (patience of 3 epochs on validation loss) and the remaining eight development participants for training, the model never saw the held-out test partition during training or early-stopping selection. Training halted after 7 epochs by the early-stopping criterion. This CNN is explicitly designated an exploratory comparator throughout the study: by prespecified protocol, it cannot replace the frozen classical primary model irrespective of its measured test-set performance, because its role is to characterise the size of the potential accuracy gap attributable to learned raw-signal representations, not to serve as a candidate for deployment-oriented selection within this study's frozen-model framework.

3.9 Reproducibility and Leakage Controls

Every stage of the pipeline was executed under a fixed global random seed (20260822), applied to NumPy, Python's hash seed, and, where applicable, TensorFlow's deterministic-operations setting. A structured leakage audit was generated programmatically at the end of the pipeline, recording: the development and test participant identifiers and confirming zero overlap between them, confirmation that participant identity was never used as a predictor, confirmation that no synthetic data, oversampling, or generative augmentation was used at any stage, confirmation that the raw ECG signal was never used as a predictor (only the deposited, ECG-derived reference HR was used, and solely as the regression target), and a record of the selection basis for the

primary model (development out-of-fold participant-macro MAE) together with explicit confirmation that the test partition was not used for selection. Software and library versions (Python, NumPy, pandas, SciPy, scikit-learn, and, where used, TensorFlow) were recorded for provenance. All intermediate artefacts — source manifests, per-participant integrity records, the engineered feature table, split definitions, out-of-fold predictions, test-set predictions, bootstrap distributions, permutation-importance tables, and figures — were retained and archived with SHA-256 checksums to support independent verification of the reported results.

4. Results

4.1 Dataset characteristics and engineered features

After preprocessing and quality screening, 64,697 eight-second windows were retained across all fifteen participants (Table 4), from an initial pool constrained by the shorter of the BVP- and ACC-derived alignment limits for each participant, no participant contributed fewer than 1,000 aligned windows, and the retained sample spans reference heart rates from 41.7 to 187.0 bpm, confirming that both resting and vigorous-exercise physiology are well represented. Table 1 summarises the eighteen engineered features across the full retained sample. The dominant PPG spectral peak (ppg_peak1_hz) had a mean of 1.22 Hz (SD 0.44), corresponding to a mean instantaneous HR estimate in the same range as the deposited reference (89.4 bpm mean), while the PPG peak-to-median power ratio showed a very large standard deviation (mean 74.2, SD 135.6, maximum 5,349), indicating that windows vary enormously in how cleanly a single spectral peak dominates the PPG spectrum — precisely the property that motion-robust estimation must accommodate. The mean absolute gap between the dominant PPG and dominant motion spectral peaks was 0.75 Hz (SD 0.63 Hz), confirming that in a substantial subset of windows the two spectra compete for overlapping frequency content.

The wide dispersion of the PPG peak-to-median power ratio and the correspondingly wide range of the PPG–motion peak gap together illustrate why a fixed, non-adaptive spectral-peak rule cannot be expected to perform uniformly well across this dataset: in low-motion windows the PPG spectrum is typically dominated by a single, sharp, high-contrast peak that a naïve rule identifies correctly, whereas in high-motion windows the PPG spectrum frequently contains multiple comparably sized peaks — one plausibly cardiac, others plausibly motion-related — for which a rule based solely on the single largest peak has no principled way to distinguish the physiologically correct candidate. This is precisely the ambiguity that the engineered secondary- and tertiary-peak features, spectral entropy, and peak-to-median ratio are designed to resolve when supplied to a learned regressor, and their prominence in the permutation-importance analysis (Section 4.5) is consistent with this design rationale.

4.2 Development-Only Model Selection

Table 3 reports the five-fold, participant-grouped, out-of-fold cross-validation performance of the three candidate regression models within the development partition.

Table 1. Descriptive statistics of the eighteen engineered spectral and time-domain features (full retained sample, N = 64,697 windows)

Feature	Mean	SD	Min	Max
ppg_peak1_hz	1.216	0.444	0.500	4.000
ppg_peak2_hz	1.216	0.467	0.500	4.000
ppg_peak3_hz	1.323	0.651	0.500	4.000
ppg_centroid_hz	1.434	0.279	0.641	2.788
ppg_entropy	2.297	0.400	1.041	3.322
ppg_peak_ratio	74.201	135.641	1.804	5349.219
motion_peak_hz	1.251	0.916	0.500	4.000
motion_centroid_hz	1.787	0.402	0.000	3.379
motion_entropy	2.808	0.264	−0.000	3.367
motion_peak_ratio	11.908	14.001	0.000	524.065
ppg_motion_peak_gap_hz	0.751	0.629	0.000	3.500
ppg_sd	1.000	0.000	1.000	1.000
ppg_iqr	1.255	0.304	0.064	1.885
ppg_diff_rms	0.317	0.064	0.130	0.702
motion_mean	1.452	0.206	0.000	1.730
motion_sd	0.874	0.238	0.000	1.710
motion_iqr	0.855	0.314	0.000	2.462
motion_diff_rms	0.641	0.323	0.000	2.431

Table 2. Participant-level development / held-out test partition distribution

Partition	Participants (n)	Windows	Mean HR (bpm)	SD HR (bpm)	Participant IDs
Development	10	43,032	89.35	21.23	S3, S4, S6, S7, S10, S11, S12, S13, S14, S15
Held-out test	5	21,665	89.58	25.70	S1, S2, S5, S8, S9

Table 3. Development-Only, Out-Of-Fold (5-Fold, Participant-Grouped) Model-Selection Results

Model	OOF participant-macro MAE (bpm)	OOF pooled MAE (bpm)	OOF RMSE (bpm)	OOF R^2
Extra-Trees (frozen primary)	8.804	8.445	13.508	0.595
Histogram Gradient Boosting	9.063	8.708	13.671	0.585
Ridge regression	13.378	12.757	17.112	0.350

Table 4. Dataset and signal-alignment summary

Metric	Value
Dataset	PPG-DaLiA (UCI Machine Learning Repository, ID 495)
Dataset DOI	10.24432/C53890
License	CC BY 4.0
Total participants	15 (S1–S15)
Archive size / integrity	2,865,111,320 bytes, SHA-256 verified
Wrist PPG (BVP) native sampling rate	64 Hz
Wrist tri-axial accelerometer sampling rate	32 Hz
Common analysis sampling rate	32 Hz (BVP downsampled by pairwise averaging)
Window length / shift	8 s / 2 s
Window sequence shape	256 samples × 4 channels (1 PPG + 3 acceleration axes)
Total retained windows (all participants)	64,697
Reference HR range (retained sample)	41.7–187.0 bpm

The Extra-Trees regressor achieved the lowest out-of-fold participant-macro MAE (8.80 bpm), narrowly outperforming histogram-based gradient boosting (9.06 bpm) and substantially outperforming ridge regression (13.38 bpm). The corresponding pooled out-of-fold MAE values were 8.45, 8.71, and 12.76 bpm respectively, and out-of-fold R^2 was 0.595 for Extra-Trees, 0.585 for gradient boosting, and 0.350 for ridge regression. The substantial gap between the linear model and both tree ensembles indicates that nonlinear interactions among the engineered spectral and motion descriptors — for example, between the PPG–motion peak gap and the PPG spectral entropy — contribute materially to predictive accuracy, consistent with the physiological expectation that the reliability of a spectral-peak-based estimate depends jointly on several interacting signal-quality indicators rather than on any single descriptor in isolation. Following the prespecified selection rule, the Extra-Trees regressor was designated the primary model, refit on the entire development partition, and frozen prior to any exposure to the test partition.

4.3 Held-Out Test-Partition Evaluation

Table 5 reports point estimates for both the frozen primary (Extra-Trees) model and the deterministic spectral baseline on the untouched test partition (five participants, 21,665 windows), formal uncertainty quantification for the difference between the two approaches is reported separately in Table 6 via paired, participant-clustered bootstrap contrasts. The primary model achieved a test MAE of 14.12 bpm and RMSE of 19.71 bpm, compared with 24.61 bpm

and 37.79 bpm respectively for the spectral baseline. Pearson correlation between the primary model's estimates and the ECG-derived reference was 0.646 (versus 0.242 for the baseline), and the concordance correlation coefficient — which more strictly penalises both imprecision and systematic bias — was 0.567 for the primary model versus 0.193 for the baseline. R^2 was 0.412 for the primary model, while the spectral baseline produced a negative R^2 (−1.161), indicating that the deterministic peak-picking approach performed worse than simply predicting the mean reference HR for every test window, a direct quantitative illustration of how severely uncompensated motion artefact degrades naïve spectral HR estimation on this closer-to-real-life activity protocol. Median absolute error, which is less sensitive to the heavy-tailed, motion-driven outliers evident in the Bland–Altman analysis below, was more similar between the two approaches (10.25 bpm for the primary model versus 11.78 bpm for the baseline), indicating that the primary model's advantage over the baseline is concentrated substantially — though not exclusively — in its performance on the more difficult, higher-error windows. Mean absolute percentage error followed the same pattern as MAE and RMSE (15.70% for the primary model versus 24.76% for the baseline), and, because MAPE is scale-normalised by the reference HR itself, this consistency across both absolute and relative error metrics indicates that the primary model's advantage is not an artefact of the particular HR range represented in the test partition, but reflects a genuine proportional improvement in estimation accuracy across the observed physiological range.

A more nuanced pattern emerges from the accuracy-within-threshold metrics. At the wider 10-bpm tolerance, the primary model and spectral baseline performed almost identically (49.3% versus 48.4% of test windows within 10 bpm respectively), a similarity confirmed statistically in Table 6 below. At the tighter 5-bpm tolerance, however, the spectral baseline was, unexpectedly, the better-performing approach (38.7% of windows within 5 bpm for the baseline versus 30.9% for the primary model). Considered alongside the baseline's much larger MAE, RMSE, and negative R^2, this pattern indicates that the deterministic spectral-peak rule follows a distinctly bimodal error distribution: it is frequently very close to the reference HR in low-motion windows, where the dominant PPG spectral peak genuinely is the cardiac peak, but catastrophically wrong in high-motion windows, where the spectral-peak rule instead locks onto the motion frequency. The primary model, by contrast, appears to trade away some of this baseline's best-case sharpness in exchange for substantially more consistent, moderate performance across the full range of window difficulty — a smoothing behaviour typical of ensemble-averaged regression relative to a discrete, rule-based decision — which the aggregate MAE, RMSE, and CCC metrics summarise as a clear overall improvement despite the baseline's local advantage at the narrowest tolerance.

Table 6 reports the paired, participant-clustered bootstrap contrasts. The primary model's MAE was 10.49 bpm lower than the baseline's (95% CI −18.41 to −4.02 bpm), and its RMSE was 18.08 bpm lower (95% CI −26.53 to −8.10 bpm), both intervals exclude zero, indicating a statistically clear improvement over the deterministic baseline once participant-level clustering is accounted for. The within-10-bpm accuracy proportion, by contrast, showed no clear difference between the two approaches (primary 49.3% versus baseline 48.4% of test windows within 10 bpm of the reference, paired difference 0.9 percentage points, 95% CI −3.3 to 5.8 percentage points, an interval that includes zero). This apparent tension — a large, statistically clear improvement in MAE and RMSE alongside no clear improvement in the proportion of windows within a fixed 10-bpm threshold — indicates that the primary model's advantage is driven disproportionately by a reduction in the magnitude of the largest errors (which RMSE and MAE are sensitive to) rather than by a uniform shift of many borderline windows across the 10-bpm threshold, this pattern is visible directly in the Bland–Altman analysis below, where the spectral baseline exhibits markedly wider limits of agreement than the primary model.

Table 5. Held-out test-partition point metrics for the frozen primary model and the deterministic spectral baseline.

Model	MAE (bpm)	RMSE (bpm)	Median AE (bpm)	Pearson r	CCC	R^2	MAPE (%)	Within 5 bpm (%)	Within 10 bpm (%)
Primary (Extra-Trees)	14.120	19.715	10.248	0.646	0.567	0.412	15.70	30.87	49.28
Spectral baseline	24.605	37.791	11.782	0.242	0.193	−1.161	24.76	38.70	48.38

Table 6. Paired, participant-clustered bootstrap contrasts (primary model minus spectral baseline, 2,000 resamples, 95% CI).

Contrast (primary − spectral baseline)	Estimate	95% CI (low)	95% CI (high)
MAE difference (bpm)	−10.485	−18.413	−4.019
RMSE difference (bpm)	−18.076	−26.530	−8.099
Within-10-bpm accuracy difference (percentage points)	0.90	−3.32	5.81

4.4 Agreement Analysis and Error Distribution

Figure 2 presents the hexbin density scatter of the primary model's estimates against the ECG-derived reference HR (left panel) and the corresponding Bland–Altman plot (right panel) for the full test partition. The scatter shows dense agreement concentrated around the identity line for reference HR in the resting-to-moderate range (roughly 60–100 bpm), which comprises the majority of the test-partition windows, with a visible fan of overestimation at higher reference HR (above approximately 130 bpm) where vigorous-activity windows are concentrated and motion artefact is most severe. The Bland–Altman plot shows a small positive mean bias (approximately 2 bpm) with 95% limits of agreement spanning roughly −38 to +42 bpm, the width of these limits — driven by a minority of high-error windows rather than the bulk of the distribution — quantifies the residual difficulty of the small subset of high-motion windows even after the feature-based motion-context information is incorporated. The asymmetric spread visible in the Bland–Altman plot, with a heavier tail of positive differences (overestimation) at higher mean HR, is consistent with periods of intense physical activity in which the motion spectral peak can itself be mistaken for the cardiac peak.

Figure 3 decomposes test-partition MAE by participant for both the primary model and the spectral baseline. The primary model outperformed the baseline for every one of the five test participants (S1, S2, S5, S8, S9), but the magnitude of both models' errors varied substantially across participants: participant S5 showed the highest error for both approaches (primary model approximately 23 bpm, baseline approximately 47 bpm), while participants S1 and S2 showed the lowest error for both approaches (primary model below 11 bpm). This participant-level heterogeneity — consistent with the pattern reported by Wilkosz and Szczęsna for a deep-learning architecture on the same dataset — indicates that a meaningful share of the residual error variance is attributable to participant-specific factors (such as individual activity intensity, sensor-skin coupling, or physiological range) rather than to model architecture alone, and it further motivates the use of participant-clustered rather than window-pooled uncertainty quantification throughout this study.

4.5 Feature Importance

Figure 4 reports held-out permutation feature importance for the frozen primary model. The dominant PPG spectral peak frequency (ppg_peak1_hz) was overwhelmingly the most influential feature, with permutation producing an increase in test MAE of approximately 4.2 bpm — roughly double the reliance placed on any other single feature.

The secondary PPG spectral peak (ppg_peak2_hz, increase of approximately 1.8 bpm), PPG spectral entropy (ppg_entropy, approximately 1.3 bpm), and the root-mean-square of the PPG first difference (ppg_diff_rms, approximately 1.1 bpm) formed a second tier of moderately influential features, followed by the PPG spectral centroid and the third PPG spectral peak. Accelerometer-derived features individually contributed comparatively modest permutation importance (each below approximately 0.3 bpm), despite the accelerometer channel's central conceptual role in motion characterisation, this pattern is consistent with the interpretation that the model's motion-awareness is substantially encoded through interaction effects — for example, the joint configuration of the PPG spectral shape (entropy, peak prominence, secondary-peak structure) together with the PPG–motion peak gap — rather than through the accelerometer's marginal, single-feature predictive contribution, a nuance that permutation importance, as a single-feature attribution method, can only partially capture.

4.6 Exploratory deep-learning comparator

The exploratory one-dimensional CNN, trained directly on raw normalised window sequences and evaluated under the identical test partition, achieved a test MAE of 10.93 bpm, RMSE of 16.86 bpm, median absolute error of 6.16 bpm, Pearson correlation of 0.758, CCC of 0.717, R^2 of 0.570, MAPE of 12.00%, 44.4% of windows within 5 bpm, and 63.9% of windows within 10 bpm of the reference (Table 7) — a consistent improvement over the frozen classical primary model across every reported metric, including, notably, the tighter 5-bpm threshold at which the classical primary model had trailed even the spectral baseline (Section 4.3).

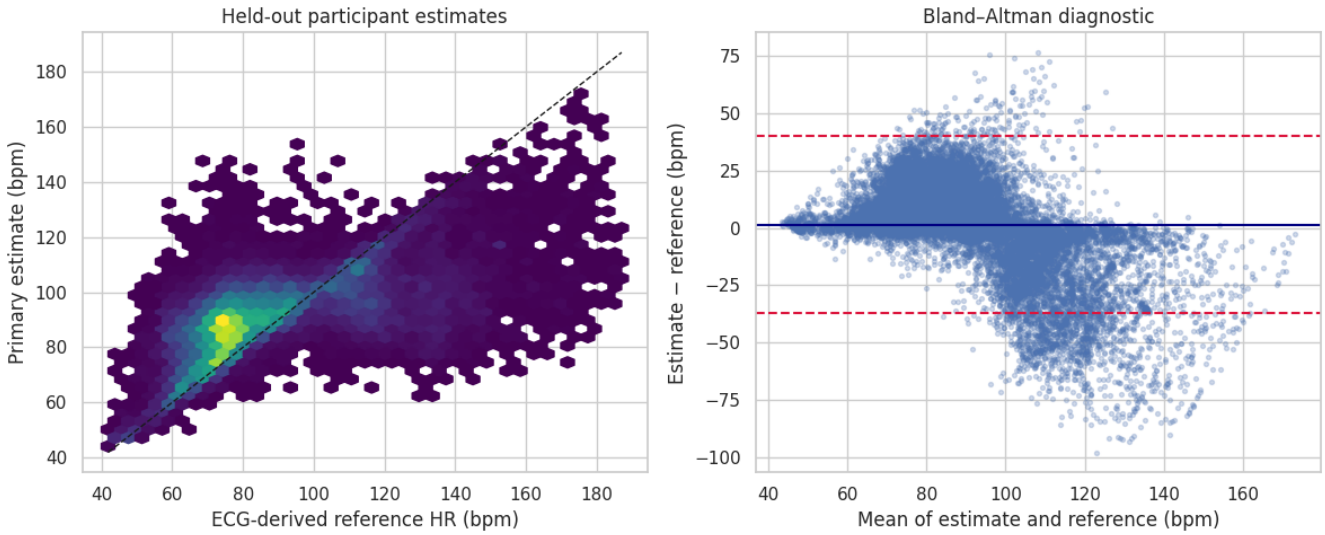

Figure 2. Held-out test-partition agreement between the frozen primary (Extra-Trees) model and ECG-derived reference heart rate. Left: hexbin density scatter of primary-model estimates versus reference HR, with the identity line shown dashed. Right: Bland–Altman diagnostic plot (estimate minus reference against the mean of the two), with the mean bias (solid line) and 95% limits of agreement (dashed lines).

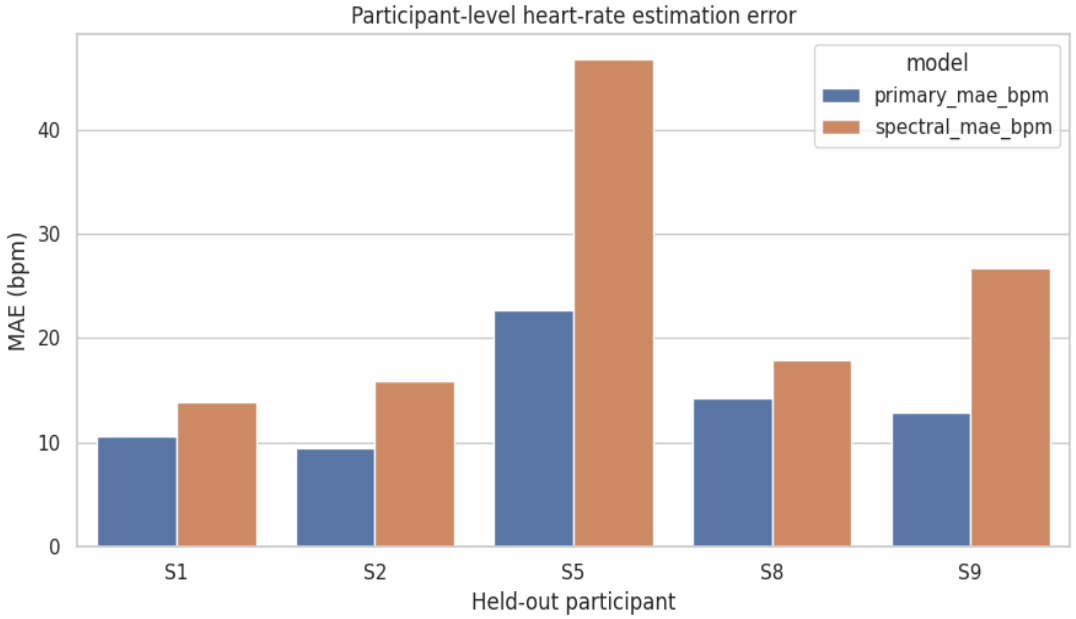

Figure 3. Per-participant test-partition mean absolute error (MAE) for the frozen primary (Extra-Trees) model and the deterministic spectral baseline, for each of the five held-out test participants (S1, S2, S5, S8, S9).

This last point is informative: unlike the primary model, which appeared to trade best-case sharpness for consistency relative to the baseline, the CNN comparator improved simultaneously on both the aggregate error metrics and the tightest accuracy-within-threshold metric, suggesting that learning directly from the raw waveform allows the network to resolve at least some of the ambiguous, high-motion windows correctly rather than merely averaging over them. This result indicates that a compact, end-to-end representation-learning approach can recover additional accuracy beyond that captured by our eighteen hand-engineered spectral and time-domain descriptors, plausibly by learning temporal or morphological structure in the raw PPG waveform (such as pulse-shape distortion under motion) that is not fully summarised by the frequency-domain and dispersion statistics used here. Consistent with the prespecified protocol described in Section 3.7, this exploratory result does not alter the frozen status of the classical Extra-Trees model as the primary output of this study, we discuss its implications for future work in Section 5.4.

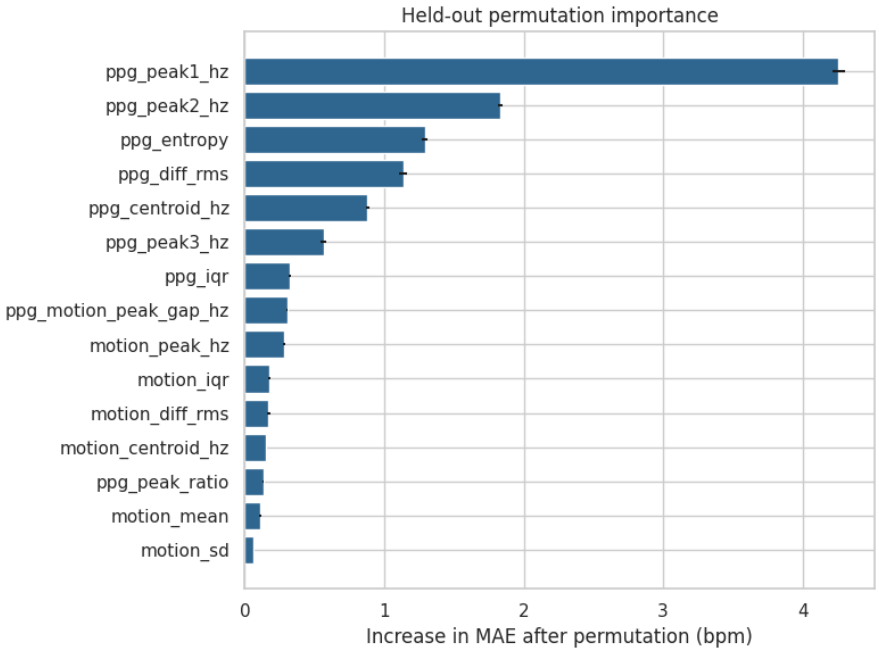

Figure 4. Held-out permutation feature importance for the frozen primary (Extra-Trees) model, expressed as the mean increase in test-set MAE (bpm) when each feature is independently permuted (12 repeats per feature).

Table 7. Exploratory 1-D CNN comparator: held-out test-partition metrics.

Metric	Value
MAE (bpm)	10.935
RMSE (bpm)	16.858
Median absolute error (bpm)	6.160
Pearson r	0.758
Concordance correlation coefficient (CCC)	0.717
R^2	0.570
MAPE (%)	12.00
Within 5 bpm (%)	44.39
Within 10 bpm (%)	63.91
Internal validation participants	S14, S15
Training epochs run (early stopping)	7

4.7 Reproducibility and Leakage Audit

The programmatic leakage audit confirmed zero overlap between the development and test participant sets, confirmed that participant identity was never used as a predictor, and confirmed that no synthetic data, oversampling, or data augmentation was applied at any stage of the pipeline. The audit further confirmed that the raw ECG signal was never used as a model input and that the deposited ECG-derived reference HR was used exclusively as the regression target. All fourteen prespecified output artefacts (integrity manifests, feature tables, split definitions, model-selection tables, bootstrap distributions, test predictions, permutation-importance tables, and figures) were generated successfully, and the complete result archive was checksummed (SHA-256) for provenance.

5. Discussion

5.1 Principal Findings

This study developed a motion-aware, feature-based Extra-Trees regressor for wrist-PPG heart-rate estimation and evaluated it with a level of participant-independent, statistically calibrated rigour that is not uniformly present in the existing PPG-DaLiA literature. The frozen primary model reduced test-partition MAE by 10.49 bpm relative to a deterministic spectral-peak baseline (95% CI −18.41 to −4.02 bpm) and more than tripled the concordance correlation coefficient (0.567 versus 0.193), demonstrating that a modest set of eighteen interpretable spectral and time-domain descriptors, combined with a classical tree-ensemble regressor, captures a substantial share of the motion-robustness that more elaborate signal-decomposition frameworks pursue through explicit algorithmic design. At the same time, the persistently wide Bland–Altman

limits of agreement (approximately −38 to +42 bpm) and the pronounced participant-to-participant heterogeneity in error (Figure 3) indicate that this accuracy is neither uniform across individuals nor sufficient, without further development, for applications demanding tight per-window accuracy during vigorous activity.

5.2 Interpreting the Feature-Importance Profile and the Mechanism of Motion Robustness

The dominance of the primary PPG spectral peak frequency in the permutation-importance analysis (Figure 4, approximately 4.2 bpm increase in MAE upon permutation) confirms the physiologically expected result that the location of the dominant PPG spectral peak carries most of the direct information about instantaneous HR — precisely the quantity the deterministic spectral baseline also relies upon exclusively. The critical difference between the primary model and the baseline is therefore not that the model discovers an entirely different information source, but that it learns, from data, when the naïve spectral-peak reading is trustworthy and when it is not, using the secondary PPG peak, spectral entropy, peak-to-median ratio, and the PPG–motion frequency gap as contextual signal-quality indicators. This interpretation is consistent with the relatively modest individual permutation importance of the accelerometer-derived features: their contribution to accuracy appears to operate largely through interaction with PPG-derived descriptors (for instance, corroborating or contradicting the dominant PPG peak) rather than through independent, additive predictive power, a pattern that single-feature permutation importance can only partially reveal and that motivates future work using interaction-aware or Shapley-based attribution methods on this feature set.

5.3 Comparison with Prior Work on PPG-Dalia

The frozen primary model's test MAE of 14.12 bpm is higher than the participant-averaged MAE figures reported for tuned deep-learning architectures on PPG-DaLiA in the literature — for example, 6.28 bpm reported by Wilkosz and Szczęsna for their multi-headed Conv-LSTM architecture, and the roughly 31% MAE reduction relative to an extended classical baseline reported by Reiss et al. for their original CNN. Several methodological differences likely contribute to this gap and should temper any direct comparison. First, published deep-learning benchmarks on PPG-DaLiA frequently employ substantially more elaborate architectures, longer and more carefully tuned training schedules, and, in some cases, post-hoc smoothing of the HR trajectory across consecutive windows (exploiting the fact that true HR cannot change arbitrarily fast), none of which were applied here in order to preserve a clean, per-window, methodologically transparent evaluation. Second, participant-split composition varies across studies and is not always identical, and — as documented in Section 4.3 — test-partition difficulty in this dataset is highly participant-dependent, so headline MAE figures computed on different five-participant test partitions are not strictly comparable without matched splits. Third, and most directly informative, our own exploratory CNN comparator (Section 4.6), trained under the identical partitioning, preprocessing, and evaluation protocol used for the primary model, achieved a test MAE of 10.93 bpm — meaningfully better than the classical primary model but still above the best published figures, which is consistent with our comparator's deliberately compact architecture (developed as an exploratory check rather than a tuned, publication-optimised deep model) and supports the conclusion that the residual gap between our results and the strongest published deep-learning benchmarks reflects architectural and training-protocol maturity rather than a fundamental limitation of learning from this dataset under strict participant-independent evaluation.

A further, more subtle comparison concerns how error is distributed rather than merely its central tendency. Wilkosz and Szczęsna explicitly reported that their Conv-LSTM architecture's error was highly participant-dependent, with one participant carrying a disproportionate share of total error, our own results show the same qualitative pattern (Section 4.4, Figure 3), with participant S5 contributing roughly double the error of the best-performing test participants under both the primary model and the spectral baseline. The persistence of this participant-dependent error structure across two independently developed model families — a Conv-LSTM network in the cited study and a feature-based Extra-Trees ensemble together with a compact CNN here — suggests that at least part of the residual, hardest-to-eliminate error in PPG-DaLiA is attributable to genuine, participant-specific factors (such as unusually intense or atypical activity execution, individual differences in peripheral perfusion, or transient sensor-skin coupling issues) rather than to any single modelling choice, and that averaged, participant-macro accuracy figures reported in isolation — without an accompanying account of the underlying participant-level spread — risk overstating the reliability of any single-number performance claim in this literature.

5.4 Classical Ensembles versus Deep Representation Learning: An Accuracy–Auditability Trade-Off

The approximately 3-bpm MAE gap between the frozen classical primary model and the exploratory CNN comparator, observed under matched partitioning, quantifies a concrete accuracy cost of choosing an engineered-feature, classical-ensemble

pipeline over an end-to-end deep architecture on this dataset. Whether that cost is acceptable depends on the deployment context. The Extra-Trees pipeline requires only eighteen closed-form spectral and statistical descriptors, is trainable and interpretable with commodity tooling, and its permutation-importance profile provides a direct, auditable account of which signal properties drive its predictions — a property directly relevant to regulatory and clinical-adjacent contexts in which model transparency is often a prerequisite for adoption. The compact CNN, by contrast, achieves higher accuracy but at the cost of substantially reduced interpretability and a training and deployment pipeline that is more demanding to audit, version, and validate on embedded wearable hardware, echoing the resource-constrained trade-offs explicitly discussed by Reiss et al. in their own resource-constrained CNN variant. We view the two approaches as complementary rather than competing: the classical pipeline established here provides an interpretable, computationally frugal baseline and a diagnostic tool for understanding which signal properties matter, while the exploratory comparator quantifies the achievable accuracy ceiling that future, more carefully engineered deep architectures — evaluated under the same participant-independent discipline — might approach without sacrificing methodological rigour.

5.4.1 Practical implications for wearable deployment

For practitioners deciding between these two families of approach in a real deployment, three deployment-relevant considerations follow directly from our results. First, the classical Extra-Trees pipeline requires no specialised inference hardware: the eighteen engineered features are computable with closed-form spectral and statistical operations well within the computational budget of contemporary wearable microcontrollers, and the trained ensemble itself can be evaluated with a small, fixed number of decision-tree traversals per prediction, making it directly suitable for on-device, real-time inference without a dedicated accelerator. Second, because every feature has a specific, human-interpretable signal-processing meaning, the permutation-importance profile reported in Section 4.5 provides engineers and regulators with a direct, auditable explanation of the circumstances under which the model's predictions should be trusted less — specifically, windows with high spectral entropy, closely spaced or comparably sized PPG peaks, or a small PPG–motion peak gap — which can be operationalised as a real-time confidence or signal-quality flag alongside the HR estimate itself, a capability that is considerably harder to construct transparently for an end-to-end deep network. Third, the approximately 3-bpm accuracy advantage of the exploratory CNN comparator indicates that applications requiring the tightest achievable accuracy — for example, precise cardiovascular research instrumentation — may still warrant the additional engineering and validation investment of a deep architecture, provided that investment includes the same participant-independent evaluation discipline demonstrated here, rather than assuming that architectural sophistication alone guarantees genuine generalisation.

5.5 The role of participant-independent evaluation and calibrated uncertainty

A secondary contribution of this study is methodological rather than purely empirical: the explicit, auditable separation between model development (restricted entirely to a ten-participant development partition with participant-grouped cross-validation) and a single, terminal evaluation on five untouched participants, combined with participant-clustered bootstrap uncertainty quantification throughout. The practical consequence of this discipline is visible directly in the results: had accuracy been assessed via simple window-level cross-validation without participant grouping, or reported as pooled point estimates without clustering-aware confidence intervals, the reported performance would very plausibly have appeared both higher and more precise than the estimates in Table 5, because window-level splits allow information about a specific participant's physiology and sensor-skin coupling to leak between training and evaluation, and unclustered confidence intervals understate uncertainty when thousands of dependent windows are drawn from only a handful of independent participants. We suggest that participant-clustered evaluation and uncertainty quantification of this kind should be treated as a minimum methodological standard for wearable-sensor HR-estimation studies intending to make generalisation claims, consistent with the broader emphasis on rigorous validation articulated in the 2023 wearable-photoplethysmography roadmap.

6. Limitations

Several limitations qualify the interpretation of these results. First, the held-out test partition comprised only five participants, while this is consistent with, and indeed larger than, several prior PPG-DaLiA evaluation protocols, it necessarily limits the precision of participant-clustered inference, as reflected in the comparatively wide bootstrap confidence intervals reported in Table 5 and Table 6. Second, the fixed random seed used for the development/test participant split, while methodologically defensible and fully documented, represents a single realisation of the fifteen-participant partition space, repeated, multiply-randomised participant splits (a form of nested or repeated grouped cross-validation across the full

participant pool) would provide a more complete characterisation of split-dependent variability, at correspondingly greater computational cost, and constitutes a natural extension of this work. Third, the eighteen engineered features, while physiologically motivated, represent one reasonable but non-exhaustive characterisation of the PPG and accelerometer spectra, alternative descriptors (for example, wavelet-based decompositions, higher-order spectral moments, or morphological pulse-shape features) might further improve accuracy or interpretability and were outside the scope of the present, deliberately parsimonious feature set. Fourth, hyperparameters for all three candidate regression models were fixed a priori with light manual tuning rather than subjected to an exhaustive, nested hyperparameter search within the development partition, while this choice preserves methodological transparency and avoids any risk of implicit overfitting to the development folds through extensive search, it may understate the accuracy achievable by each candidate architecture under more extensive tuning. Fifth, the exploratory CNN comparator was deliberately compact and was not subjected to architecture search or extensive training-schedule tuning, so its reported performance should be interpreted as a lower bound on what representation-learning approaches can achieve on this dataset under our partitioning protocol, not as a ceiling. Finally, PPG-DaLiA itself, though larger and more activity-diverse than earlier benchmarks, comprises only fifteen participants recorded with a single wrist-worn device model under a single research protocol, generalisation to other devices, sensor placements, skin tones, age ranges, and free-living (rather than protocol-driven) activity contexts cannot be established from this dataset alone and would require external validation, consistent with broader calls in the wearable-photoplethysmography literature for cross-device and cross-population validation.

7. Conclusion

This study developed and rigorously evaluated a motion-aware, feature-based machine-learning pipeline for wrist-PPG heart-rate estimation on the PPG-DaLiA dataset, using a fully prespecified, participant-independent evaluation protocol with calibrated uncertainty quantification throughout. A frozen Extra-Trees regressor, trained on eighteen interpretable spectral and time-domain descriptors of synchronised PPG and tri-axial accelerometry, reduced held-out test MAE by 10.49 bpm relative to a deterministic spectral-peak baseline (95% CI −18.41 to −4.02 bpm) and more than tripled the concordance correlation coefficient, while remaining fully interpretable through permutation feature importance, which identified the dominant and secondary PPG spectral peaks and PPG spectral entropy as the leading predictors. An exploratory convolutional neural network, evaluated under an identical protocol but explicitly barred from replacing the frozen classical model, achieved further accuracy gains, quantifying a concrete — though not overwhelming — advantage of raw-signal representation learning over engineered features on this task. Taken together, these findings indicate that lightweight, auditable, classical ensemble methods remain a competitive and practically attractive option for motion-robust wearable heart-rate estimation, particularly where interpretability, computational economy, and methodological transparency are prioritised alongside accuracy, and they underscore that participant-independent evaluation with clustered uncertainty quantification — rather than model architecture alone — is essential to producing trustworthy, generalisable performance claims in this domain. Future work should extend this protocol to repeated, multiply-randomised participant splits, richer or learned feature representations, and external validation across additional devices and populations, in order to establish the robustness of these findings beyond the specific dataset and partition studied here. Three extensions appear particularly promising given the findings reported above. First, because the accelerometer-derived features individually showed comparatively low permutation importance despite their conceptual centrality to motion characterisation (Section 4.5), interaction-aware attribution methods (for example, Shapley-value decompositions applied to the same frozen model) could clarify whether their contribution operates chiefly through interaction with PPG-derived descriptors, refining the feature set accordingly. Second, given the participant-dependent error structure observed here and corroborated in independent prior work, a systematic study explicitly modelling participant-level random effects — rather than treating participants solely as a clustering unit for uncertainty quantification — could help distinguish genuinely irreducible physiological heterogeneity from heterogeneity attributable to correctable modelling limitations. Third, the roughly 3-bpm gap between the classical primary model and the exploratory CNN comparator, both evaluated under an identical protocol, motivates a follow-up study in which a more carefully tuned deep architecture is developed and evaluated under the same frozen-model, participant-independent discipline established here, to determine how much of that gap reflects genuine representational advantage versus the specific architectural and training choices made for the present exploratory comparator.

References

[1] P.H. Charlton, J. Allen, R. Bailón, S. Baker, J.A. Behar, F. Chen, G.D. Clifford, D.A. Clifton, H.J. Davies, C. Ding, X. Ding, J. Dunn, M. Elgendi, M. Ferdoushi, D. Franklin, E. Gil, M.F. Hassan, J. Hernesniemi, X. Hu, N. Ji, Y. Khan,

S. Kontaxis, I. Korhonen, P.A. Kyriacou, P. Laguna, J. Lázaro, C. Lee, J. Levy, Y. Li, C. Liu, J. Liu, L. Lu, D.P. Mandic, V. Marozas, E. Mejía-Mejía, R. Mukkamala, M. Nitzan, T. Pereira, C.C.Y. Poon, J.C. Ramella-Roman, H. Saarinen, M.M.H. Shandhi, H. Shin, G. Stansby, T. Tamura, A. Vehkaoja, W.K. Wang, Y.T. Zhang, N. Zhao, D. Zheng, T. Zhu, The 2023 wearable photoplethysmography roadmap. Physiological Measurement, 44(11), (2023) 111001. [DOI]

[2] B. Bent, B.A. Goldstein, W.A. Kibbe, J.P. Dunn. Investigating sources of inaccuracy in wearable optical heart rate sensors. NPJ Digital Medicine, 3, (2020) 18. [DOI]

[3] Z. Zhang, Z. Pi, B. Liu, TROIKA: A general framework for heart rate monitoring using wrist-type photoplethysmographic signals during intensive physical exercise. IEEE Transactions on Biomedical Engineering, 62(2), (2015) 522–531. [DOI]

[4] Z. Zhang, Photoplethysmography-based heart rate monitoring in physical activities via joint sparse spectrum reconstruction. IEEE Transactions on Biomedical Engineering, 62(8), (2015) 1902–1910. [DOI]

[5] A. Reiss, I. Indlekofer, P. Schmidt, & K. Van Laerhoven. (2019). Deep PPG: Large-scale heart rate estimation with convolutional neural networks. Sensors, 19(14), Article 3079. [DOI]

[6] A. Reiss, I. Indlekofer, P. Schmidt. (2019). PPG-DaLiA [Dataset]. UCI Machine Learning Repository.

[7] M. Wilkosz, A. Szczęsna. Multi-headed Conv-LSTM network for heart rate estimation during daily living activities. Sensors, 21(15), (2021) 5212. [DOI]

[8] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees. Machine Learning, 63(1), (2006) 3–42. [DOI]

[9] J.H. Friedman, Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), (2001) 1189–1232. [DOI]

[10] J.M. Bland, D.G. Altman, Statistical methods for assessing agreement between two methods of clinical measurement. The Lancet, 327(8476), (1986) 307–310. [DOI]

[11] L.I.K. Lin. A concordance correlation coefficient to evaluate reproducibility. Biometrics, 45(1), (1989) 255–268. [DOI]

[12] L. Breiman. Random forests. Machine Learning, 45(1), (2001) 5–32. [DOI]

[13] A. Fisher, C. Rudin, F. Dominici, All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. Journal of Machine Learning Research, 20(177), (2019) 1–81.

[14] B. Efron, Bootstrap methods: Another look at the jackknife. The Annals of Statistics, 7(1), (1979) 1–26. [DOI]

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(85), (2011) 2825–2830.

Funding

No funding was received for the conduct of this research

Conflict of interest

The authors declare that no conflict of interest exists.

Does the Article Screened for Similarity?

Yes.

About the License

© The Author(s) 2026. The text of this article is open access and licensed under a Creative Commons Attribution 4.0 International License