Proceedings of ICBioMed 2026 Symposium: Biomedical Engineering and Public Health: Real-Time Wearables for Disease Detection
DOI: 10.54254/2753-8818/2026.ES37117

Beyond Average Accuracy: Motion-Related Failure Patterns in Wrist-Worn Photoplethysmographic Heart-Rate Monitoring

Angel Xu

The Leys School, Cambridge, UK
axu961224@gmail.com

Abstract. Wrist-worn photoplethysmography (PPG) enables unobtrusive heart rate monitoring, but motion-related errors tend to cluster under specific conditions and are often masked by aggregate accuracy metrics. This study characterized those failure patterns and assessed whether wrist-sensor features could identify unreliable estimates in previously unseen participants. PPG-DaLiA recordings from 15 participants, acquired with an Empatica E4 and synchronized chest electrocardiography during daily-life activities, were segmented into 8-s windows with a 2-s step. Heart rate was estimated using a transparent dominant-frequency PPG method. A large error was defined a priori as an absolute deviation of more than 10 beats per minute (bpm) relative to the ECG-derived reference. Motion intensity, acceleration-band power, PPG-acceleration spectral overlap, peak dominance and spectral entropy were used as reliability predictors. Of 47,182 non-transition windows, 14,622 (31.0%) met the large-error criterion. Failure prevalence differed substantially by activity, from 5.8% during sitting to 63.8% during stair climbing. In leave-one-subject-out validation, a motion-only model achieved an AUROC of 0.743, wrist-signal features increased AUROC to 0.851, and activity context increased it to 0.864. Wrist PPG reliability is therefore condition dependent, and an interpretable signal-based reliability flag may complement point heart-rate estimates. External prospective validation is required before clinical or consumer deployment.

Keywords: wearable sensors, motion artefact, heart-rate monitoring, signal quality

1. Introduction

Photoplethysmography (PPG) estimates pulsatile blood-volume changes optically and is now embedded in wrist-worn devices for continuous heart-rate monitoring. Its low burden makes it valuable for ambulatory research, exercise feedback and remote health monitoring. However, the optical waveform is measured at a moving, deformable interface. Changes in contact pressure, arm movement and tissue deformation can create spectral components that compete with the cardiac pulse. Validation studies have therefore reported substantially different errors across activities and individuals, even within the same device class [1-3].

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2026.ES37117

Most algorithmic studies approach this task by minimizing average heart rate estimation error or maximizing overall benchmark performance. Classical approaches use frequency-domain peak selection, sparse spectral reconstruction or motion-adaptive filtering [4-6], while more recent work uses convolutional and recurrent neural networks [7, 8]. These methods improve estimation, but a single mean absolute error does not answer an operational question: when should a displayed reading be treated as unreliable? Signal-quality models have begun to address this issue [9, 10]. Neural estimation in daily living settings [11], ecologically valid within-person validation [12], and foundational analysis of PPG waveform morphology [13] further demonstrate that algorithmic design, contextual factors, and measurement conditions must be considered jointly.

The PPG-DaLiA dataset was created to represent daily-living conditions and provides synchronized wrist PPG, wrist acceleration, activity labels and ECG-derived reference heart rate [14]. It therefore permits a focused analysis of failure rather than another optimization contest. The present study asks: (1) which activities and motion conditions are associated with large errors from a transparent wrist-PPG estimator, (2) how much do failure rates vary between participants, and (3) can wrist-signal features prospectively flag large-error windows for a held-out participant? The study's contribution is an interpretable, reproducible description of condition-dependent failures and a leakage-resistant reliability benchmark. The analysis is associational and does not interpret a 10-bpm error as a clinical safety threshold.

2. Materials and methods

2.1. Dataset and analysis unit

PPG-DaLiA was obtained from the UCI Machine Learning Repository (DOI: 10.24432/C53890) [14]. Fifteen adults performed sitting, stair climbing, table soccer, cycling, driving, lunch, walking and working, with additional periods coded as transitions. The Empatica E4 supplied wrist BVP at 64 Hz and tri-axial acceleration at 32 Hz. A chest-worn RespiBAN supplied ECG at 700 Hz. The synchronized SX.pkl files contain corrected R-peak indices and an ECG-derived label for each 8-s segment shifted by 2 s, the label is the mean instantaneous heart rate in that segment [14].

The primary analysis used windows assigned to one of the eight named activities. Windows coded as transition activity 0 were excluded because they combine movement between tasks and do not represent a stable activity condition. A sensitivity analysis retaining transition windows is specified in the prespecified protocol. S6 was retained with its available valid recording, the dataset documentation reports a hardware interruption for this participant.

2.2. Transparent PPG estimator

For each 8-s BVP window, the signal was detrended and filtered using a fourth-order zero-phase Butterworth band-pass filter (0.7-3.5 Hz, equivalent to 42-210 beats per minute). A Hann taper and zero-padded periodogram were used to obtain a frequency spacing of approximately 0.01 Hz. The largest spectral peak within the passband was selected and converted to beats per minute. No ECG value, participant-specific calibration, activity label or future window was used by the estimator. This deliberately simple estimator renders failure mechanisms interpretable and provides a fixed target for reliability analysis, rather than representing an optimized, state-of-the-art accuracy benchmark.

DOI: 10.54254/2753-8818/2026.ES37117

2.3. Motion and signal-quality features

Acceleration signals were mean-centered within each window prior to calculation of the Euclidean norm of the acceleration vector. The primary motion measure was the root-mean-square of this dynamic acceleration magnitude. Additional features were acceleration-magnitude standard deviation and acceleration power in the 0.7-3.5-Hz cardiac band. To quantify potential motion contamination, spectral overlap was defined as acceleration power within 0.10 Hz of the PPG-selected peak, divided by total acceleration power in the cardiac band. PPG peak dominance was the selected peak's power divided by total in-band PPG power. Normalized Shannon entropy of the in-band PPG spectrum was used as a dispersion measure. All features were calculated without ECG information.

The feature set was selected to represent three complementary mechanisms. Motion RMS measures movement amplitude but does not indicate whether movement occupies the cardiac frequency band. Acceleration band power and spectral overlap address frequency-specific competition. PPG peak dominance and spectral entropy describe whether the optical signal itself contains a concentrated, interpretable cardiac component. This separation is important because two windows can have similar acceleration RMS while differing in whether the acceleration spectrum collides with the selected PPG peak. Conversely, a low-motion window can still be unreliable if optical contact is poor or the pulsatile component is weak. The features are therefore not a complete physiological model, they form a compact and interpretable representation available to a wrist device at the time of estimation.

Frequency resolution also affects interpretation. An 8-s window contains a limited number of cardiac cycles, particularly at low heart rates. Zero padding provides a finer grid for peak localization but does not create new information, it interpolates the discrete spectrum. The estimate should therefore be understood as the dominant frequency supported by the window, not as a beat-by-beat measurement. This limitation is one reason the study focuses on reliability classification and error distributions rather than presenting the estimator as a clinical-grade monitor.

2.4. Outcomes and quality control

The signed error was the PPG estimate minus the ECG reference, absolute error was its absolute value. The primary outcome was a large-error window (absolute error >10 bpm), prespecified as an analytical reliability threshold. Secondary outcomes were mean, median, 90th- and 95th-percentile absolute error, signed bias, proportions above 5 and 20 bpm, and consecutive large-error episodes. Missing or nonfinite windows were excluded. Reference values below 40 or above 210 bpm were excluded as physiologically implausible, with counts recorded by participant. No window was removed because it had a large error.

2.5. Statistical analysis

Error distributions were summarized by activity and participant. For adjusted associations, a population-averaged logistic generalized estimating equation (GEE) with participant as the clustering unit estimated the odds of a large-error window. Predictors were log-transformed where positively skewed and standardized using the full descriptive analysis scale: motion RMS, acceleration cardiac-band power, spectral overlap, peak dominance and spectral entropy, together with activity indicators. Results are reported as odds ratios (ORs) with 95% confidence intervals. These estimates describe associations and do not establish that motion causes error.

DOI: 10.54254/2753-8818/2026.ES37117

For window i from participant j, Y(i,j) equals 1 when absolute error exceeds 10 bpm and 0 otherwise. The population-averaged association model was specified as:

$$logit\{Pr[Y(i,j)=1]\}=\beta_0+\beta_1X_1(i,j)+\beta_2X_2(i,j)+\beta_3X_3(i,j) \quad (1)$$

$$+\beta_4X_4(i,j)+\beta_5X_5(i,j)+\Sigma(k=2,...,8)\gamma_kA_k(i,j).$$

Here, X_1 through X_5 are the standardized values of log-transformed motion RMS, log-transformed acceleration cardiac-band power, log-transformed spectral overlap, PPG peak dominance and PPG spectral entropy, respectively. $A_k(i,j)$ is an activity indicator, sitting is the reference activity. The GEE uses participant clustering to obtain robust standard errors under repeated, correlated windows. Exponentiating a coefficient gives the odds ratio for a one-standard-deviation increase in its predictor while holding the remaining measured variables fixed.

2.6. Reliability prediction and validation

Four models were compared: M0, a training-fold prevalence baseline, M1, logistic regression using motion RMS, M2, regularized logistic regression using all wrist-signal features, and M3, the same features plus activity. Evaluation used leave-one-subject-out (LOSO) cross-validation. Standardization, encoding, fitting and the decision threshold were performed within each training fold. The threshold was chosen by the maximum Youden index in the training data and then applied once to the held-out participant. Primary metrics were AUROC, area under the precision-recall curve (AUPRC), Brier score, sensitivity, specificity, positive predictive value and negative predictive value. Pooled out-of-fold predictions and participant-level metric distributions were retained.

The model hierarchy has an engineering interpretation. M0 represents a system that knows only training-set prevalence and cannot adapt to the current signal. M1 asks whether one computationally inexpensive motion statistic is sufficient. M2 tests whether signal quality and frequency-domain information add value without requiring a manually assigned activity. M3 represents the most context-aware version. Comparing adjacent models quantifies the incremental value of signal quality and context instead of rewarding a larger model simply for having more parameters.

The reliability task was separated from heart-rate estimation. The ECG reference defined the retrospective outcome but was never provided to the predictor. In deployment, ECG would not be available, the model would operate on the wrist signals used to calculate the displayed estimate. This separation prevents circularity and makes the model output a warning about signal conditions rather than a second estimate of heart rate.

For each predictive model, the estimated probability was $Pr[Y(i,j) = 1/\{1 + \exp[-\eta(i, j)]\}$, where $\eta(i,j)$ is a linear combination of training-fold standardized features. The L2-regularized objective for M2 and M3 was the negative Bernoulli log-likelihood plus $\lambda||\beta||^2$. This penalty constrains unstable coefficients when correlated frequency-domain features coexist, while retaining a direct mapping from each feature to a reliability probability. The regularization strength and all transformations were fixed within the training data of each LOSO fold.

Calibration was considered alongside discrimination. A high AUROC can coexist with systematically overconfident probabilities, particularly when event rates differ between activities or users. The Brier score therefore supplied a proper scoring rule that penalizes both ranking errors and probability miscalibration. Sensitivity and specificity were evaluated at a threshold selected inside each training fold. No test-set threshold was chosen after observing held-out outcomes. Positive

DOI: 10.54254/2753-8818/2026.ES37117

predictive value was interpreted together with event prevalence because its practical meaning changes when large-error windows are rare or common.

2.7. Reproducibility and implementation

Processing was implemented in Python using NumPy, SciPy, pandas, scikit-learn and statsmodels. Raw files were treated as immutable, and their folder and signal structures were audited before processing. The processing script generated one row per window and retained the subject identifier, window index, activity, reference and estimated heart rate, error, failure indicator and predictor features. Analysis scripts generated summary tables, failure episodes, GEE estimates and out-of-fold predictions. Figures were created from machine-readable tables rather than manually transcribed values. The protocol, scripts and generated outputs were stored together so that numerical claims can be reproduced without hidden spreadsheet operations.

3. Results

3.1. Activity-dependent error patterns

The final analysis contained 47,182 windows from 15 participants, 14,622 windows (31.0%) exceeded the 10-bpm threshold. Activity-specific results are shown in Table 1 and Figure 1. Sitting had the lowest mean absolute error (2.66 bpm) and a 5.8% failure rate. Working (6.23 bpm, 16.2%) and driving (9.42 bpm, 25.4%) were intermediate. Cycling, walking, table soccer and stair climbing had much larger tail errors. Stair climbing produced the largest MAE (33.18 bpm), 90th-percentile absolute error of 80.20 bpm and a 63.8% failure rate. Walking and table soccer had failure rates of 59.1% and 62.2%, respectively, despite differing in nominal activity type. Thus, activity labels reveal patterns that are not represented by a single resting-versus-exercise contrast.

Table 1. Error distribution by activity

Activity	Windows	MAE (bpm)	Median AE (bpm)	90th percentile AE (bpm)	Large-error windows (%)
Sitting	4,569	2.66	0.89	5.27	5.8
Stairs	3,239	33.18	26.34	80.20	63.8
Table soccer	2,310	23.10	17.47	57.64	62.2
Cycling	3,473	23.25	2.20	78.00	39.0
Driving	6,843	9.42	3.34	27.90	25.4
Lunch	13,554	9.77	3.33	30.15	26.6
Walking	4,697	24.43	18.48	59.55	59.1
Working	8,497	6.23	1.77	19.05	16.2

DOI: 10.54254/2753-8818/2026.ES37117

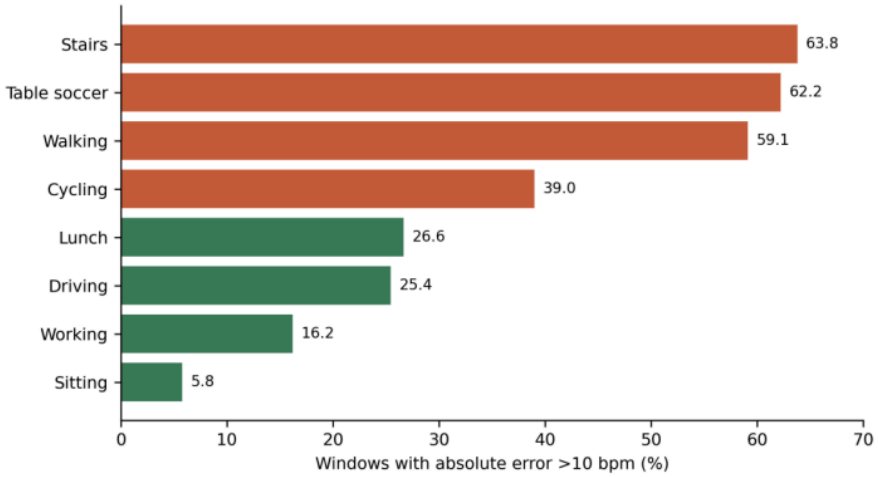


Figure 1. Proportion of windows with absolute heart-rate error greater than 10 bpm by activity

3.2. Participant heterogeneity

Participant failure rates varied from 15.1% (S13) to 51.3% (S9), while MAE ranged from 7.08 bpm (S7) to 31.20 bpm (S5) (Figure 2). The variation persisted even though every participant used the same sensor configuration and estimator. Because wear tightness, skin contact, physiology and activity execution are not separately identified in this dataset, participant differences are interpreted as heterogeneity in observed failure patterns rather than a ranking of device quality.

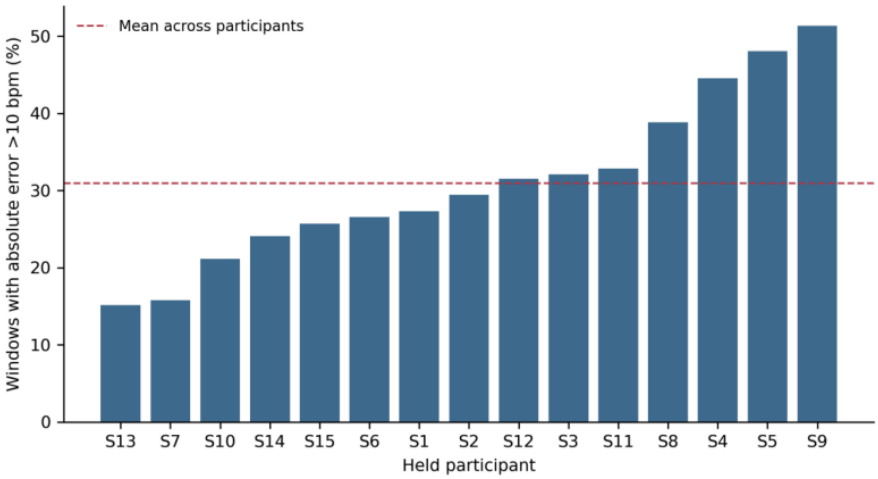


Figure 2. Between-participant variation in the proportion of large-error windows

3.3. Adjusted associations

In the participant-clustered GEE, a one-standard-deviation increase in log motion RMS was associated with higher odds of large error (OR 1.34, 95% CI 1.27-1.41). Spectral overlap was also positively associated with failure (OR 1.52, 95% CI 1.35-1.71), whereas greater PPG peak dominance (OR 0.24, 95% CI 0.20-0.28) and lower spectral entropy (OR 0.55, 95% CI 0.46-0.65) were associated with lower odds. Relative to sitting, stairs had the highest adjusted activity association (OR 4.38, 95% CI 2.77-6.92), followed by walking (OR 3.77, 95% CI 2.26-6.30) and

DOI: 10.54254/2753-8818/2026.ES37117

table soccer (OR 3.13, 95% CI 2.05-4.77). These results are descriptive of the measured data and should not be read causally.

3.4. Held-out reliability prediction

LOSO results are summarized in Table 2 and Figure 3. Motion RMS alone (M1) achieved AUROC 0.743, AUPRC 0.500 and Brier score 0.188. Adding wrist-signal features (M2) increased AUROC to 0.851 and reduced the Brier score to 0.145. Adding activity context (M3) provided a further, smaller increase to AUROC 0.864 and reduced the Brier score to 0.140. At training-fold Youden thresholds, M3 achieved 86.7% sensitivity, 72.3% specificity, 58.4% positive predictive value and 92.3% negative predictive value on pooled held-out windows. The baseline was not treated as a meaningful classifier, its purpose was to show the performance gain over a prevalence-only rule.

Table 2. Pooled leave-one-subject-out reliability performance

Model	Inputs	AUROC	AUPRC	Brier score	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
M1	Motion RMS	0.743	0.500	0.188	83.3	56.8	46.4	88.3
M2	Wrist signal features	0.851	0.664	0.145	85.9	69.9	56.2	91.7
M3	Wrist features and activity	0.864	0.683	0.140	86.7	72.3	58.4	92.3

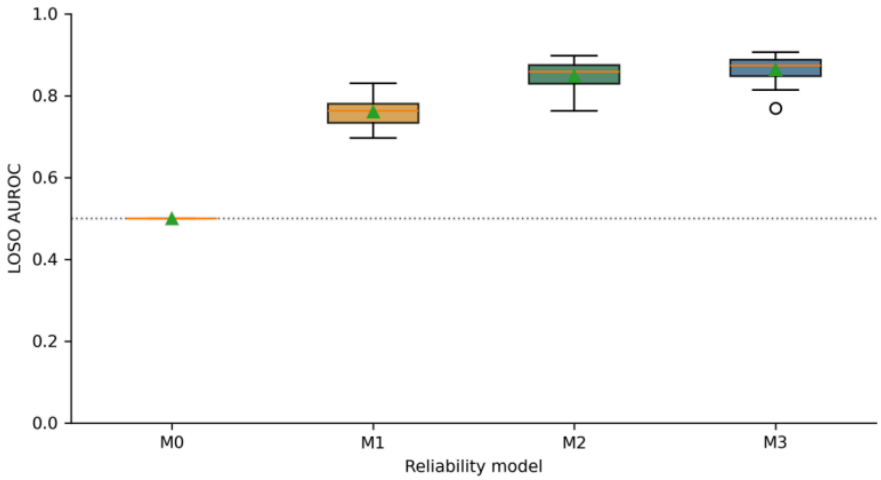


Figure 3. Distribution of participant-level AUROC values under leave-one-subject-out validation

The hierarchy indicates where predictive information arose. The improvement from M1 to M2 was larger than the improvement from M2 to M3. Motion amplitude captured a substantial risk gradient, but PPG peak structure and frequency overlap added information about whether the optical waveform supported the current estimate. Activity context remained informative after these features were included, although its incremental contribution was smaller. This pattern supports an architecture in which signal-derived quality metrics form the primary reliability layer and activity recognition supplies supplementary context.

DOI: 10.54254/2753-8818/2026.ES37117

The held-out metrics provide evidence of transportability within this dataset, not a validated clinical rule. Each test participant contributed no training windows, scaling statistics or threshold selection. The results therefore measure relative risk discrimination in a new participant under the same recording protocol. They do not establish performance with a new sensor, different sampling rates, a different population or another definition of unacceptable error. Nevertheless, this design is a stronger generalization test than a random window split.

4. Discussion

This study moves the evaluation of wrist-worn PPG beyond average accuracy toward conditional reliability. Three findings are central. First, large errors were strongly activity dependent: only 5.8% of sitting windows exceeded 10 bpm, compared with 63.8% during stair climbing and more than 59% during walking and table soccer. Second, failure prevalence varied substantially between participants, despite a common sensor configuration and estimator. Third, wrist-derived signal features identified unreliable readings in participants excluded from model fitting. Together, these findings show that one aggregate MAE conceals when, and for whom, a heart-rate estimate is most likely to fail.

The activity pattern was not equivalent to a simple sedentary-versus-exercise contrast. Cycling involved substantial movement but had a lower failure rate than stairs, walking or table soccer. Wrist posture, arm swing, grip, vibration and the frequency structure of movement may therefore matter more than motion amplitude alone. This interpretation is supported by the adjusted associations: motion RMS and PPG-acceleration spectral overlap were associated with higher failure odds, whereas a dominant PPG peak and a less dispersed PPG spectrum were associated with lower odds. These observations are consistent with motion components competing with the cardiac peak, as described in motion-robust spectral approaches [4-6]. They remain associational, however, because activity, heart rate, posture and contact pressure were not experimentally separated.

The LOSO analysis provides the main engineering result. Motion intensity alone achieved an AUROC of 0.743, while the full wrist-signal model reached 0.851, adding activity context increased AUROC to 0.864. The larger improvement from M1 to M2 indicates that frequency overlap and PPG peak structure contain information beyond the amount of movement. The smaller gain from M2 to M3 suggests that activity recognition is useful but not essential to the reliability layer. Participant-level separation is important because overlapping windows from one recording are highly correlated. Randomly assigning such windows to training and test sets could allow a model to recognize participant-specific signal properties rather than general failure conditions. LOSO validation provides a more credible assessment of transfer to a new participant, although it is still internal validation within one dataset.

Reliability prediction should complement, rather than replace, heart-rate estimation. A wearable system could report both a point estimate and a confidence state based on contemporaneous wrist signals. A low-confidence state might prompt remeasurement, delay an automated alert or mark a value for cautious interpretation. This is particularly relevant when errors occur in sustained episodes: an isolated outlier may be suppressed by a robust smoother, whereas consecutive failures can create a misleading trend. Reporting median and upper-tail errors, failure prevalence and episode duration therefore provides information that MAE alone cannot supply. The 10-bpm cutoff used here is an analytical boundary for studying this distribution, not a clinical safety threshold.

Between-participant heterogeneity further limits the meaning of pooled performance. Failure prevalence ranged from 15.1% to 51.3%, and participant-level MAE ranged from 7.08 to 31.20 bpm. These differences cannot be attributed to one mechanism because the dataset does not

DOI: 10.54254/2753-8818/2026.ES37117

independently measure strap tightness, sensor contact, wrist geometry or how each activity was performed. They nevertheless show why average performance should be accompanied by participant-level results. A model that performs well overall may remain unreliable for a subset of users. Repeated sessions and direct contact-quality measures would be required to distinguish stable individual susceptibility from session-specific wear conditions.

Several limitations define the scope of the findings. The dominant-frequency estimator was deliberately transparent and was not optimized to represent a commercial device or a state-of-the-art neural algorithm. The observed error distribution is therefore an interpretable benchmark, not a universal estimate of consumer-wearable performance. PPG-DaLiA contains 15 adults following one broadly standardized protocol, and S6 has a shorter valid recording because of a documented hardware interruption. Generalization to other devices, age groups, skin characteristics and unconstrained environments remains uncertain. The ECG labels are averages of instantaneous heart rate within 8-s windows rather than beat-level clinical adjudications. Overlapping windows also share most of their samples, participant-clustered GEE and LOSO validation reduce, but do not eliminate, the implications of serial dependence. Finally, centered acceleration is an incomplete representation of orientation and contact pressure, and activity indicators may capture unmeasured contextual differences. The GEE coefficients should consequently be interpreted as population-averaged associations rather than causal effects.

5. Conclusion

Average heart-rate error is an incomplete description of wrist-worn PPG performance. In a synchronized daily-living dataset, a transparent dominant-frequency estimator produced markedly different error patterns across activities and participants: only 5.8% of sitting windows exceeded 10 bpm, compared with more than 59% of walking, table-soccer and stair-climbing windows. These failures were not reducible to a simple exercise label. Objective wrist motion, spectral overlap between acceleration and the selected PPG frequency, peak dominance and spectral entropy were jointly informative about whether a reading was likely to be unreliable. A regularized reliability model evaluated on participants who were never used for fitting achieved an AUROC of 0.851 from wrist-signal features alone and 0.864 after adding activity context. The resulting contribution is therefore a failure-oriented evaluation framework and a reproducible reliability benchmark, rather than a claim of a new clinical monitor. In practical terms, a wearable should not communicate only a point estimate of heart rate, it should also communicate whether the signal conditions support confidence in that estimate. The present flag is a research prototype. Its analytical threshold, calibration and operating point require prospective validation, and the findings should not be interpreted as evidence of clinical safety or diagnostic performance. Nevertheless, the workflow demonstrates how public multimodal sensor data can support transparent, statistically disciplined assessment of when wearable biosensing becomes uncertain. This perspective may help future systems prioritize trustworthy output, targeted remeasurement and explicit uncertainty alongside raw accuracy. It also provides a concrete reporting template: reporting tail errors, activity-specific failures, between-participant variability, and held-out reliability performance rather than relying solely on one aggregate accuracy statistic.

References

[1] Reiss, A., Indlekofer, I., Schmidt, P., & Van Laerhoven, K. (2019). Deep PPG: Large-scale heart rate estimation with convolutional neural networks. *Sensors (Basel), 19*(14), 3079. doi:10.3390/s19143079.

DOI: 10.54254/2753-8818/2026.ES37117

[2] Bent, B., Goldfarb, M., Kibbe, W. A., & Dunn, J. P. (2020). Investigating sources of inaccuracy in wearable optical heart rate sensors. *NPJ Digit Med, 3*, 18. doi:10.1038/s41746-020-0226-6.

[3] Shcherbina, A., Mattsson, C. M., Waggott, D., Salisbury, H., Christle, J. W., Hastie, T., et al. (2017). Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort. *J Pers Med, 7*(2), 3. doi:10.3390/jpm7020003.

[4] Zhang, Z., Pi, Z., & Liu, B. (2015). TROIKA: A general framework for heart rate monitoring using wrist-type photoplethysmographic signals during intensive physical exercise. *IEEE Trans Biomed Eng, 62*(2), 522-531. doi:10.1109/TBME.2014.2359372.

[5] Zhang, Z. (2015). Photoplethysmography-based heart rate monitoring in physical activities via joint sparse spectrum reconstruction. *IEEE Trans Biomed Eng, 62*(8), 1902-1910. doi:10.1109/TBME.2015.2406331.

[6] Salehizadeh, S., Dao, D., Bolkhovsky, J., Cho, C., Mendelson, Y., & Chon, K. H. (2016). A novel time-varying spectral filtering algorithm for reconstruction of motion artifact corrupted heart rate signals during intense physical activities using a wearable photoplethysmogram sensor. *Sensors (Basel), 16*(1), 10. doi:10.3390/s16010010.

[7] Biswas, D., Everson, L. R., Liu, M., Panwar, M., Verhoef, B. E., Patki, S., et al. (2019). CorNET: Deep learning framework for PPG-based heart rate estimation and biometric identification in ambulant environment. *IEEE Trans Biomed Circuits Syst, 13*(2), 282-291. doi:10.1109/TBCAS.2019.2892297.

[8] Crepaldi, S. C., Wang, J., Matsumoto, F., Takeuchi, H., Watanabe, T., & Yamamoto, Y. (2025). A stability- and aggregation-based method for heart rate estimation using photoplethysmographic signals during physical activity. *Sensors (Basel), 25*(14), 4315. doi:10.3390/s25144315.

[9] Ding, C., Guo, Z., Chen, Z., Lee, R. J., Rudin, C., & Hu, X. (2024). SiamQuality: A ConvNet-based foundation model for photoplethysmography signals. *Physiol Meas, 45*(8). doi:10.1088/1361-6579/ad6747.

[10] Charlton, P. H., Argüello-Prada, E. J., Mant, J., & Kyriacou, P. A. (2025). The MSPTDfast photoplethysmography beat detection algorithm: Design, benchmarking, and open-source distribution. *Physiol Meas, 46*(3). doi:10.1088/1361-6579/adb89e.

[11] Wilkosz, M., & Szczęsna, A. (2021). Multi-headed Conv-LSTM network for heart rate estimation during daily living activities. *Sensors (Basel), 21*(15), 5212. https://doi.org/10.3390/s21155212

[12] Nelson, B. W., & Allen, N. B. (2019). Accuracy of consumer wearable heart rate measurement during an ecologically valid 24-hour period: Intraindividual validation study. *JMIR Mhealth Uhealth, 7*(3), e10828. https://doi.org/10.2196/10828

[13] Elgendi, M. (2012). On the analysis of fingertip photoplethysmogram signals. *Curr Cardiol Rev, 8*(1), 14–25. https://doi.org/10.2174/157340312801784038

[14] UCI Machine Learning Repository. (2019). *PPG-DaLiA* [dataset]. https://archive.ics.uci.edu/dataset/495/ppg+dalia. https://doi.org/10.24432/C53890