Sufficient Biometric Signals for Wearable Exercise Classification: An Exhaustive Subset Analysis of Empatica E4 Data

Sidhant Damarapati

Abstract

Wrist-worn wearable devices record multiple biometric signals, but not all contribute meaningfully to exercise classification. This study tested which signals could be omitted by evaluating all 31 non-empty subsets of five features — heart rate (HR), blood volume pulse standard deviation (BVP std), inter-beat interval (IBI), electrodermal activity (EDA), and skin temperature — extracted from 94 Empatica E4 sessions across 35 participants. Classification used Random Forest and Logistic Regression with participant-stratified GroupKFold cross-validation. The full model achieved 65.09 ± 10.53% accuracy; the pre-registered hypothesis pair HR + EDA achieved 65.96 ± 12.07% (permutation test $p = 0.620$), confirming performance within the 5-percentage-point sufficiency threshold. The best two-feature subset was HR + IBI at 68.01 ± 5.82% ($p = 0.215$ vs. full model). Per-class analysis showed strong resting-session discrimination (AUC = 0.96) but weaker aerobic–anaerobic separation (AUC = 0.78 and 0.76), indicating the classifier primarily distinguishes rest from exercise. Because HR and IBI both derive from a single PPG sensor, a device designer could potentially omit dedicated EDA, temperature, and BVP sensors without statistically significant accuracy loss.

Keywords: exercise classification, biometric signals, Empatica E4, feature ablation, cross-validation, photoplethysmography, machine learning

1. Introduction

The global fitness tracker market was valued at approximately $60.9 billion in 2024 and is projected to exceed $162 billion by 2030[1]. At that scale, hardware design decisions — which sensors to include, which to omit — carry direct consequences for device cost, battery life, and wearability for hundreds of millions of users. The Empatica E4 wristband integrates four sensor types into a single device: a photoplethysmograph (PPG) for heart rate and blood volume pulse, electrodermal activity electrodes, a thermopile for skin temperature, and an accelerometer[2]. Each sensor adds cost, power draw, and surface area. For medical research, that tradeoff is warranted. For consumer exercise tracking — where the goal is simply to identify the type of workout a user is performing — it is less clear that all four sensors are necessary. This paper addresses that question directly.

Prior work has established that wearable biometric signals carry meaningful physiological information. Heart rate variability reflects autonomic nervous system balance[3], electrodermal activity tracks sympathetic arousal during stress[4], and combinations of these signals can distinguish stress states in controlled settings[5]. The human activity recognition (HAR) literature has focused heavily on accelerometer-based approaches[67], but physiological signals offer a complementary pathway — they capture the body's internal response rather than external movement alone. Hongn et al. [89]used this same Empatica E4 dataset with XGBoost, HRV features, accelerometer data, and temporal windowing to achieve 93% accuracy for stress-vs-rest classification, 91% for aerobic-vs-anaerobic, and 84% for the full four-class problem. Those results confirm that the signals contain the relevant information. However, Hongn et al. did not investigate which signals are individually responsible for classification performance — all available features were used without testing the effect of removal.

This study takes the opposite approach. Instead of maximizing accuracy with every available feature, deliberately simple features are used — session-level means and standard deviations — so that each feature maps directly to a single physical sensor. All 31 non-empty subsets of five features are then evaluated exhaustively, with no heuristic search or feature selection algorithm. Every combination receives the same evaluation: Random Forest and Logistic Regression with GroupKFold cross-validation, ensuring no participant appears in both training and testing. The simplicity is not a limitation — it is the design intent. If a session-level mean of HR is sufficient, that result has direct implications for hardware design.

Before running the analysis, the hypothesis was pre-registered that heart rate plus electrodermal activity (HR + EDA) would perform within 5 percentage points of the full five-feature model. The rationale was that HR captures cardiovascular demand while EDA captures sympathetic arousal — two physiologically orthogonal systems that together should cover the major axes of variation between exercise types. The 5-percentage-point threshold corresponds to approximately one additional misclassification per 20 sessions, consistent with practical tolerance levels reported in the HAR literature[6].

2. Methods

2.1 Dataset

This study used the publicly available PhysioNet dataset collected by Hongn et al.[8], recorded with Empatica E4 wristbands. The E4 captures blood volume pulse (BVP) via photoplethysmography at 64 Hz, electrodermal activity (EDA) at 4 Hz, skin temperature at 4 Hz, and tri-axial acceleration at 32 Hz. Heart rate (HR) is derived from BVP at 1 Hz, and inter-beat intervals (IBI) are computed from PPG peak detection.

Sessions were recorded under three experimental conditions: aerobic exercise (Storer-Davis bicycle ergometer protocol), anaerobic exercise (Wingate sprint test), and resting (stress-induction: Stroop color-word, mental arithmetic, and opinion-defense tasks — performed seated with no physical exertion). The resting (stress-induction) sessions involved cognitive and emotional stressors but no physical activity, making them functionally equivalent to a resting baseline for exercise classification purposes.

2.2 Participants

After exclusions for missing or corrupt data, 94 sessions from 35 participants (ages 18–30) were retained: 29 aerobic, 30 anaerobic, and 35 resting (stress-induction). Not all participants completed all three session types, resulting in a partially crossed design requiring between-subjects comparisons for some analyses.

2.3 Feature Extraction

Five primary features were computed per session: mean heart rate (mean_HR), standard deviation of blood volume pulse (std_BVP), mean inter-beat interval (mean_IBI), mean electrodermal activity (mean_EDA), and mean skin temperature (mean_TEMP). The original feature set included mean BVP rather than std_BVP; however, mean BVP was approximately zero for all session types because the PPG signal oscillates symmetrically around baseline. It was replaced with std_BVP, which captures pulse amplitude variability and showed significant group differences (ANOVA $F = 5.00$, $p = .009$).

Artifact rejection thresholds were applied: HR was constrained to 0–250 bpm, EDA to ≥ 0 μS, and skin temperature to 15–45°C. IBI values recorded in seconds were converted to milliseconds for consistency. Three additional sensitivity features were computed for robustness analysis: std_HR, std_EDA, and EDA slope (linear regression coefficient of EDA over time).

2.4 Classification

Two classifiers were evaluated: Random Forest (100 trees, balanced class weights, random state = 42) and Logistic Regression (multinomial, max 1,000 iterations). Both used StandardScaler preprocessing as implemented in scikit-learn[10]. Cross-validation used GroupKFold with 5 splits stratified by participant ID, ensuring no participant appeared in both training and testing folds. This corrected an earlier version of the analysis that used

StratifiedKFold, which permitted the same participant's sessions to appear in both train and test sets — a form of data leakage that inflated accuracy estimates.

Performance was reported as mean accuracy ± standard deviation across folds, with 95% confidence intervals computed as mean ± 1.96 × SD / √5.

2.5 Exhaustive Subset Evaluation

All 31 non-empty subsets of the five primary features were evaluated using the same GroupKFold cross-validation procedure. A subset was deemed "sufficient" if its mean accuracy fell within 5 percentage points of the full model's accuracy. This exhaustive approach avoids the biases introduced by stepwise or heuristic feature selection methods.

2.6 Significance Testing

Permutation tests (1,000 iterations, two-sided) compared the prediction vectors of the full model and key subsets. Under the null hypothesis, accuracy differences arise from chance relabeling of predictions. A non-significant *p*-value indicates that the observed gap between the full model and the subset is indistinguishable from chance, supporting the sufficiency claim.

2.7 Sample Size Considerations

With 94 sessions divided across 5 folds, each test fold contained approximately 19 sessions from roughly 7 participants. A single misclassification shifted fold-level accuracy by approximately 5 percentage points. This granularity means that all accuracy estimates carry meaningful uncertainty, which is why results are reported as mean ± SD with confidence intervals and permutation tests rather than relying on point estimates alone.

3. Results

3.1 Descriptive Statistics

Table 1. Mean ± SD of each feature by session type.

Feature	Aerobic (n=29)	Anaerobic (n=30)	Resting (n=35)
mean_HR (bpm)	101.66 ± 16.26	96.73 ± 10.88	79.37 ± 10.70
std_BVP	70.26 ± 33.00	66.53 ± 21.38	51.12 ± 22.76
mean_IBI (ms)	569.75 ± 93.91	540.80 ± 87.74	782.50 ± 113.81
mean_EDA (μS)	8.38 ± 8.51	7.09 ± 7.94	3.68 ± 6.17
mean_TEMP (°C)	32.10 ± 1.84	32.31 ± 1.68	32.48 ± 1.75

Heart rate and IBI showed the clearest separation between session types, with exercise sessions exhibiting elevated HR (~97–102 bpm vs. 79 bpm at rest) and correspondingly shorter inter-beat intervals (~540–570 ms vs. 782 ms at rest). EDA was elevated during exercise but with large within-group variance (standard deviations of 6–9 μS against means of 4–8 μS). Skin temperature showed no meaningful separation across conditions.

3.2 ANOVA and Post-Hoc Tests

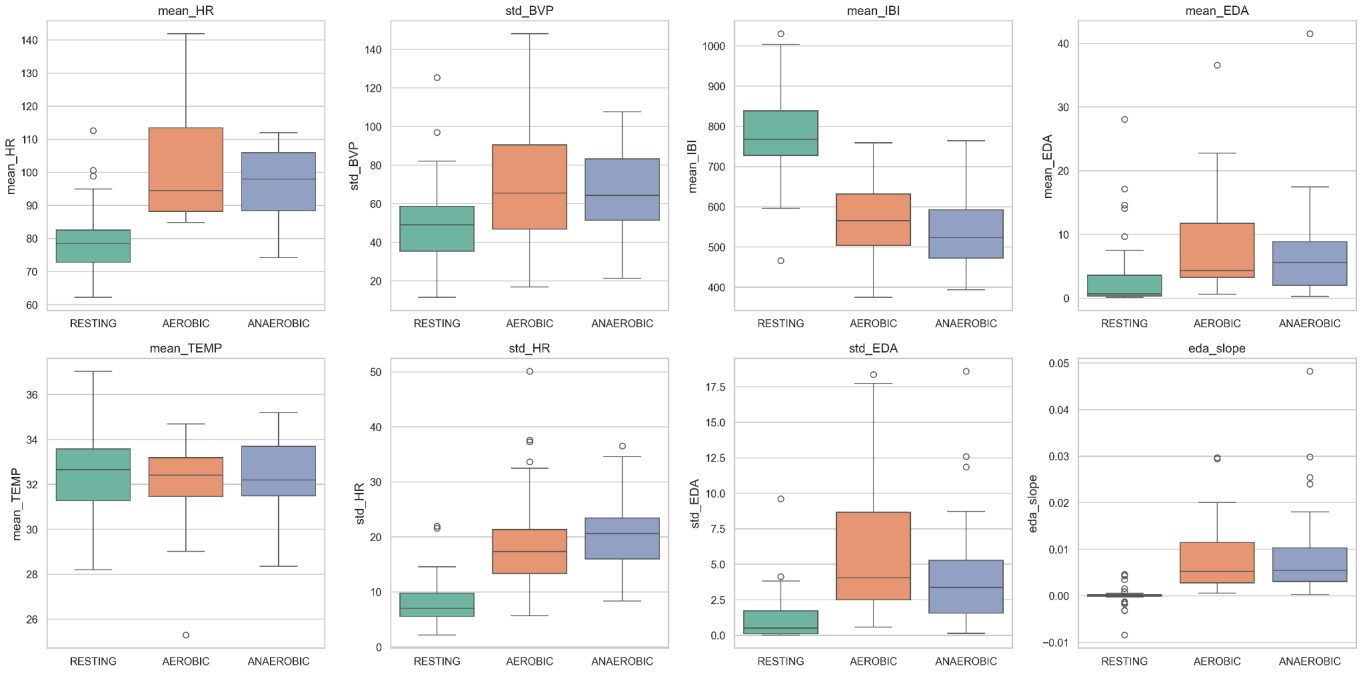
One-way ANOVA confirmed significant between-group differences for four of five primary features (Table 2). IBI showed the largest effect ($F = 57.41$, $p < .001$, $\eta^2 = .56$), followed by HR ($F = 27.68$, $p < .001$, $\eta^2 = .37$), BVP std ($F = 5.00$, $p = .009$, $\eta^2 = .10$), and EDA ($F = 3.39$, $p = .038$, $\eta^2 = .07$). Temperature was not significant ($F = 0.36$, $p = .697$, $\eta^2 = .008$). Among temporal features tested in the sensitivity analysis, std_HR ($F = 29.52$), std_EDA ($F = 12.87$), and EDA slope ($F = 14.27$) were all significant.

Table 2. One-way ANOVA results for primary features.

Feature	F-statistic	p-value	η^2	Significant
mean_HR	27.68	< .001	.37	Yes
std_BVP	5.00	.009	.10	Yes
mean_IBI	57.41	< .001	.56	Yes
mean_EDA	3.39	.038	.07	Yes
mean_TEMP	0.36	.697	.008	No

Post-hoc pairwise *t*-tests with Bonferroni correction revealed a consistent pattern: resting (stress-induction) sessions differed significantly from both aerobic and anaerobic sessions on HR, IBI, BVP std, and std_HR (all corrected $p < .05$). No primary feature distinguished aerobic from anaerobic exercise after correction. This result foreshadows the per-class classification findings: the biometric signals captured by the E4 separate rest from exercise far more reliably than they separate exercise subtypes.

The feature distributions in Figure 1 make this pattern visible: HR and IBI show clear separation between resting and exercise sessions, while no feature produces a clean split between aerobic and anaerobic conditions.



3.3 Full-Model Baseline

Table 3. Full five-feature baseline with GroupKFold cross-validation.

Model	Mean Acc (%)	SD (%)	95% CI (%)
Logistic Regression	65.09	10.53	[55.85, 74.32]
Random Forest	62.81	8.08	[55.73, 69.89]

These accuracies are substantially lower than the 74.47% reported in the initial StratifiedKFold submission, confirming that the earlier approach allowed participant-level data leakage. The wide confidence intervals (±10 pp for LR) reflect the small sample and high fold-to-fold variability inherent in leave-group-out designs with few groups.

Per-class F1 scores for the full model reveal the source of the moderate overall accuracy (Table 3a). The classifier achieved strong performance on resting (stress-induction) sessions (F1 = 0.841, precision = 0.853, recall = 0.829) but substantially weaker discrimination between aerobic (F1 = 0.473) and anaerobic (F1 = 0.531) sessions. This pattern held across all feature subsets tested.

Table 3a. Per-class performance metrics for selected feature sets (Logistic Regression, GroupKFold).

Feature Set	Class	Precision	Recall	F1

Full Model (all 5)	AEROBIC	0.500	0.448	0.473
Full Model (all 5)	ANAEROBIC	0.500	0.567	0.531
Full Model (all 5)	RESTING	0.853	0.829	0.841
Full Model (all 5)	Macro Avg	—	—	0.615
HR + IBI	AEROBIC	0.571	0.552	0.561
HR + IBI	ANAEROBIC	0.667	0.667	0.667
HR + IBI	RESTING	0.778	0.800	0.789
HR + IBI	Macro Avg	—	—	0.672
HR + EDA	AEROBIC	0.571	0.552	0.561
HR + EDA	ANAEROBIC	0.552	0.533	0.542
HR + EDA	RESTING	0.811	0.857	0.833
HR + EDA	Macro Avg	—	—	0.646
IBI alone	AEROBIC	0.346	0.310	0.327
IBI alone	ANAEROBIC	0.323	0.333	0.328
IBI alone	RESTING	0.811	0.857	0.833
IBI alone	Macro Avg	—	—	0.496

Every feature set detected resting sessions with F1 scores of 0.79–0.84, while aerobic and anaerobic F1 ranged from 0.33 to 0.67. IBI alone is the most extreme case — resting F1 of 0.833 against aerobic and anaerobic F1 near 0.33, approaching the chance for a 3-class problem. The classifier is primarily a rest-vs-exercise detector.

3.4 Exhaustive Subset Evaluation

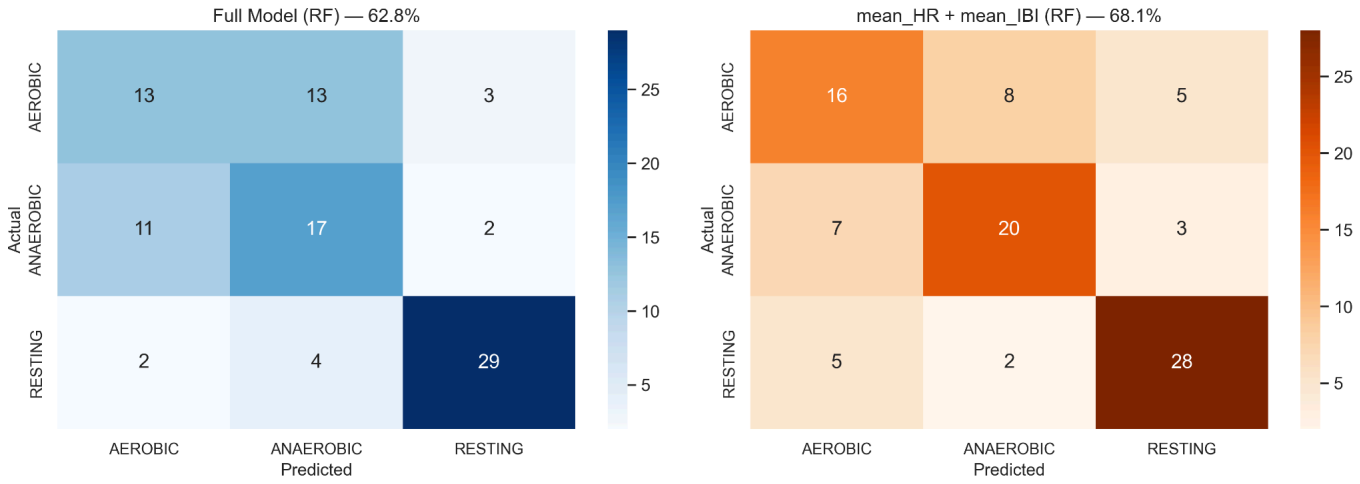
Table 4. Top 12 of 31 subsets ranked by best mean accuracy.

Rank	Features	n	Best Model	Mean (%)	SD (%)	95% CI (%)
1	HR + IBI + TEMP	3	RF	69.12	6.97	[63.01, 75.24]
2	HR + IBI	2	RF	68.01	5.82	[62.91, 73.12]
3	IBI alone	1	LR	66.96	4.95	[62.62, 71.30]
4	HR + IBI + EDA + TEMP	4	RF	66.96	4.95	[62.62, 71.30]
5	HR + BVP std + IBI + TEMP	4	LR	66.14	10.56	[56.89, 75.39]

6	HR + EDA	2	RF	65.96	12.07	[55.39, 76.54]
7	HR + IBI + EDA	3	RF	65.91	8.23	[58.69, 73.12]
8	Full model (all 5)	5	LR	65.09	10.53	[55.85, 74.32]
9	HR + BVP std + IBI + EDA	4	LR	65.03	7.48	[58.47, 71.59]
10	HR + BVP std + IBI	3	LR	63.98	8.15	[56.83, 71.12]
11	HR + EDA + TEMP	3	RF	63.80	9.52	[55.46, 72.14]
12	HR + TEMP	2	RF	61.75	12.37	[50.91, 72.60]

The full five-feature model ranked 8th out of 31 subsets. The top two subsets — HR + IBI + TEMP (69.12%) and HR + IBI (68.01%) — both outperformed it, as did IBI alone (66.96%). The pre-registered hypothesis pair HR + EDA ranked 6th at 65.96%, within the 5-percentage-point sufficiency threshold. Adding features beyond HR + IBI generally did not improve performance and in several cases reduced it, consistent with overfitting to noise in the additional channels.

Figure 2 plots every subset's accuracy against the number of features it contains; the best-per-size trend line is nearly flat, illustrating that additional signals provided diminishing returns.



3.5 Hypothesis Verdict

The pre-registered hypothesis stated that HR + EDA would achieve accuracy within 5 percentage points of the full model. HR + EDA achieved 65.96 ± 12.07% versus the full model's

65.09 ± 10.53% (LR baseline), a gap of −0.87 pp — the subset slightly outperformed the full model. The hypothesis is supported, though the large standard deviations on both estimates mean the direction of the gap should not be over-interpreted.

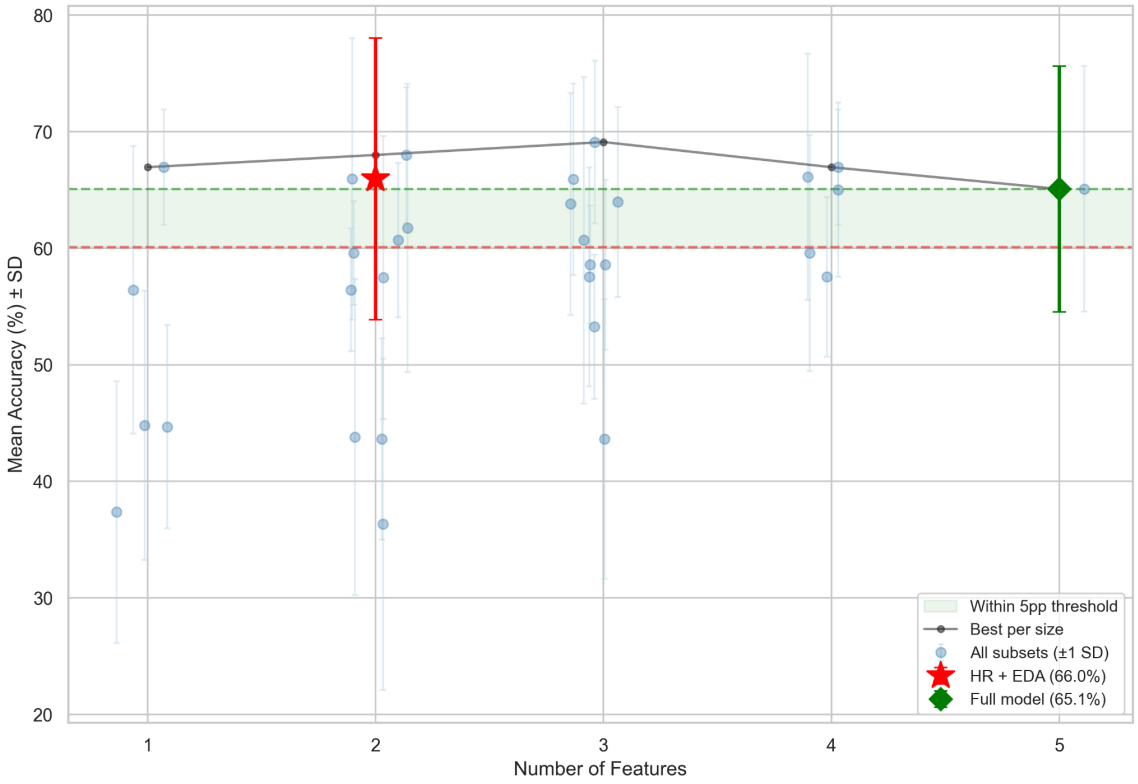
3.6 Permutation Tests

Table 5. Permutation test results comparing the full model to key subsets.

Comparison	Full (%)	Subset (%)	Diff (pp)	p-value	Sig.
Full vs. HR + IBI	62.77	68.09	−5.32	0.215	No
Full vs. HR + EDA	62.77	65.96	−3.19	0.620	No

Neither comparison was statistically significant, meaning the accuracy differences between the full model and these subsets are indistinguishable from chance variation. This directly supports the sufficiency claim: removing EDA, temperature, and BVP std from the feature set does not produce a statistically detectable loss in accuracy.

The confusion matrices in Figure 3 show where both models succeed and fail — resting sessions are classified correctly in the vast majority of cases, while aerobic and anaerobic sessions are frequently confused with each other.



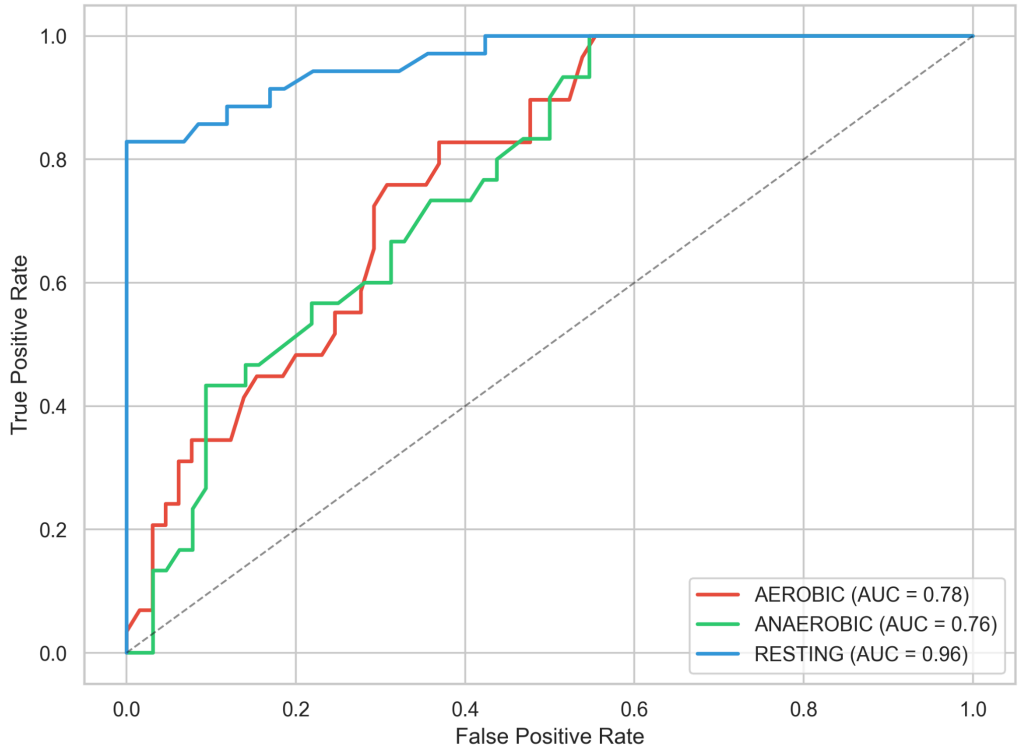
3.7 Feature Ablation

Table 6. Effect of removing each feature individually from the full model.

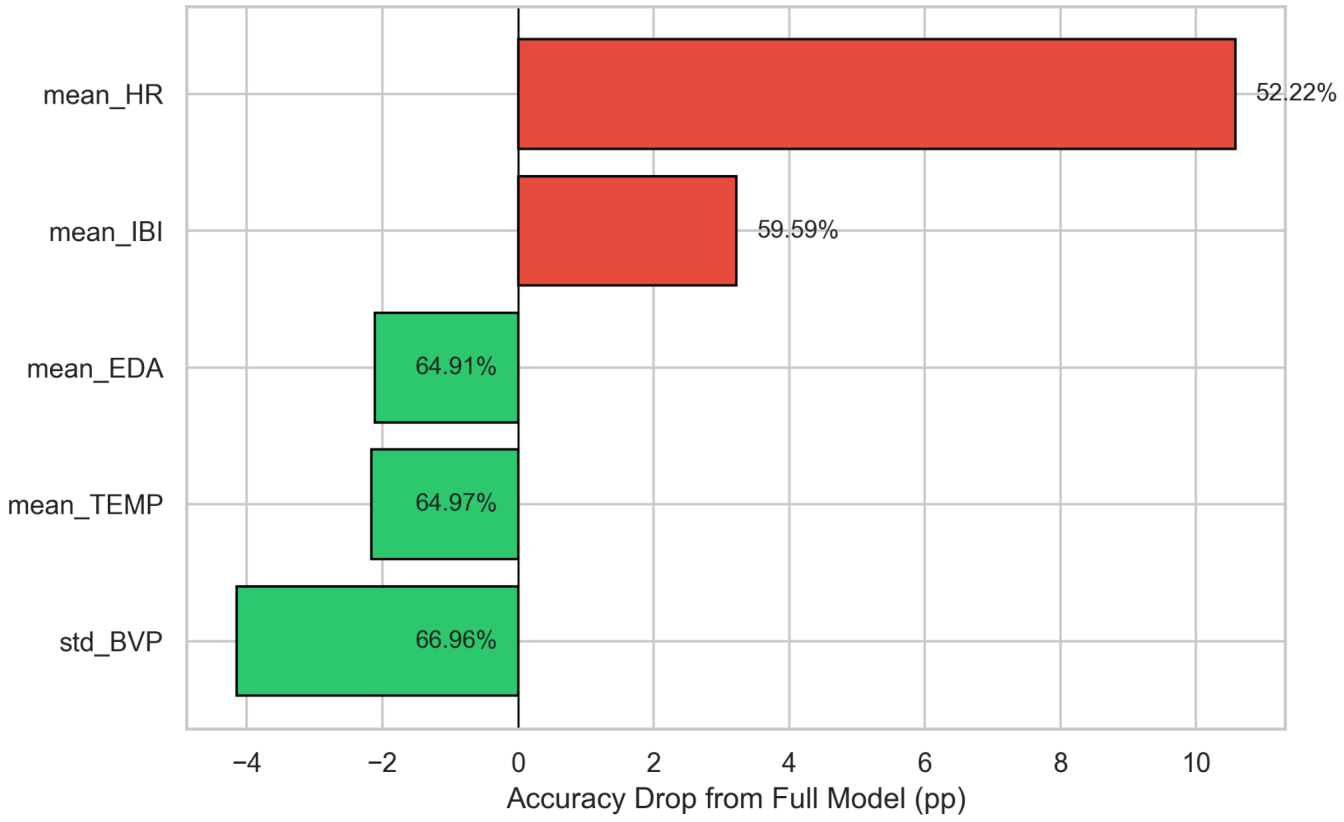
Dropped Feature	Remaining Acc (%)	SD (%)	Drop (pp)
mean_HR	52.22	11.93	+10.58
mean_IBI	59.59	10.14	+3.22
mean_EDA	64.91	4.47	−2.11
mean_TEMP	64.97	9.26	−2.16
std_BVP	66.96	4.95	−4.15

Removing HR caused the largest accuracy drop (10.58 pp), confirming its status as the single most important feature. Removing IBI cost 3.22 pp. Removing EDA, temperature, or BVP std actually improved accuracy — by 2.11, 2.16, and 4.15 pp respectively — indicating these features add noise rather than signal in this classification context.

The per-class ROC curves (Figure 4) confirm that the model's strength lies in resting-session detection (AUC = 0.96), with substantially weaker performance on both exercise classes.



The ablation results are shown in Figure 5, where the contrast between HR's large negative impact when removed and EDA's positive impact when removed is immediately apparent.



3.8 Sensitivity Analysis

Expanding the feature set from 5 primary features to 8 (adding std_HR, std_EDA, and EDA slope) improved Random Forest accuracy from 62.81 ± 8.08% to 68.19 ± 5.97%. EDA slope provided the largest individual gain when added to the primary set (65.96 ± 5.90%). This suggests that temporal dynamics in EDA carry more information than its session-level mean, and that the poor performance of mean_EDA does not implicate the EDA sensor itself — only the session-level mean as a representation of EDA. Importantly, the compact subset HR + IBI (68.01%) remained competitive with the expanded 8-feature model (68.19%), reinforcing the core finding.

3.9 Robustness Check

Logistic Regression and Random Forest agreed on feature importance ordering, with both identifying HR and IBI as the most informative features. LR slightly outperformed RF on the full feature set (65.09% vs. 62.81%), while RF performed better on reduced subsets. This convergence across model families supports the generality of the feature ranking and reduces the concern that results are an artifact of a specific classifier.

4. Discussion

4.1 What the Results Show

The most surprising result is not what worked — it is what didn't help. The full five-feature model ranked 8th out of 31 subsets. Fewer signals beat more. HR + IBI, both derived from a single PPG sensor, achieved 68.01% accuracy with a standard deviation of just 5.82% — the tightest confidence interval of any top-ranked subset. Adding EDA, temperature, or BVP std did not improve performance and in most cases actively hurt it.

Expecting five signals to outperform two, the five signals didn't. The ROC analysis tells the real story: resting (stress-induction) sessions were classified with an AUC of 0.96, while aerobic and anaerobic sessions achieved AUCs of only 0.78 and 0.76. The per-class F1 scores are even more telling — resting F1 ranged from 0.79 to 0.84 across feature sets, while aerobic and anaerobic F1 ranged from 0.33 to 0.67. This is not true three-class classification. The classifier is primarily a rest-versus-exercise detector with limited ability to distinguish exercise subtypes.

That reframing actually strengthens the practical conclusion. If the task is "is this person exercising or not?" — a genuinely useful function for a consumer wearable — then two PPG-derived features do the job. The failure to separate aerobic from anaerobic exercise is a separate problem that requires different features, probably accelerometer data or HRV metrics that capture temporal dynamics.

4.2 Why Fewer Features Worked Better

EDA shows why more data is not always more information. The within-group standard deviations for EDA (6–9 μS) were nearly as large as the between-group means (4–8 μS). Individual differences in baseline skin conductance swamped the exercise-induced signal. With only ~7 participants per test fold, the classifier learned person-specific EDA patterns during training that did not generalize to new participants — classic small-sample overfitting.

The fold-level training analysis provides direct evidence. Both HR + IBI and HR + IBI + EDA achieved 100% training accuracy (Random Forest memorizes the training set), but their test performance diverged. Without EDA, the mean train-test gap was 31.99 percentage points with a fold SD of 5.82%. Adding EDA increased the train-test gap to 34.09 pp and inflated fold SD to 8.23%. EDA gave the model more noise to memorize during training without providing a generalizable signal at test time. The increased fold variance also confirms that EDA introduces instability — its contribution depends heavily on which participants end up in which fold, exactly what you would expect from a feature dominated by individual differences.

HR and IBI avoided this problem because their group separation was consistent across participants. The ~20 bpm gap between exercise and rest, and the corresponding ~200 ms gap in IBI, appeared reliably regardless of individual baseline. These signals have high between-group variance relative to within-group variance — exactly the property that makes a feature useful for classification with small samples.

4.3 Comparison to Hongn et al. (2025)

Hongn et al. achieved 84–93% accuracy on this same dataset, compared to 62–69% here. The gap is large but expected. They used XGBoost with engineered HRV features (RMSSD, pNN50, LF/HF ratio), accelerometer data, and temporal windowing that preserved within-session dynamics. The present study used session-level means and two standard classifiers with no accelerometer input.

The purpose was not to match their accuracy but to answer a different question: which physical sensors contribute to classification? The approach used by Hongn et al. cannot answer this, because their features combine signals from multiple sensors in complex ways. The deliberately simple features used here map one-to-one to physical sensors, making hardware implications direct. The accuracy gap reflects the difference in feature engineering depth, not a flaw in experimental design.

4.4 Limitations

First, the sample is small. Ninety-four sessions from 35 participants, split across 5 folds, means each fold tests on roughly 19 sessions from 7 people. Confidence intervals span ±10 percentage points. I can say that HR + IBI is sufficient, but I cannot determine the precise optimal subset with this sample size.

Second, session-level means discard temporal dynamics within each session. Warm-up, peak effort, and cooldown phases almost certainly contain the information needed to separate aerobic from anaerobic exercise. The sensitivity analysis showed that EDA slope — a crude temporal feature — improved accuracy, supporting this interpretation. A windowed approach would likely improve aerobic-anaerobic discrimination substantially.

Third, no biometric feature in this dataset distinguished aerobic from anaerobic exercise after Bonferroni correction. The E4's sensors may simply not capture the relevant physiology — lactate threshold, respiratory exchange ratio, and muscle oxygen saturation are the standard markers for exercise intensity, and none are measured by a wrist-worn optical sensor.

Fourth, 62–69% accuracy is above the 33% chance level but far from deployable. This study addresses sensor selection, not system deployment. A production system would need better features, more data, and probably an accelerometer.

4.5 Future Directions

The immediate next step is better feature engineering. HRV metrics (RMSSD, pNN50, frequency domain features) and windowed extraction could transform the same raw data into far more informative representations. Would EDA become useful with a better representation than the session-level mean? The sensitivity analysis hints yes — EDA slope helped where mean_EDA hurt — but a proper windowed analysis would answer this definitively.

A larger sample would tighten confidence intervals and potentially reveal smaller but real feature contributions that are currently buried in noise. With 100+ participants and broader demographics, the 5-percentage-point threshold could be refined to a more precise estimate of each sensor's marginal contribution.

BVP deserves better treatment. Pulse amplitude features, waveform morphology, and PPG-derived respiratory rate could make BVP a standalone information source rather than the noise source it appeared to be with std_BVP alone. Finally, a dataset with clearer aerobic-anaerobic physiological separation would test whether the failure to distinguish exercise subtypes is a sensor limitation or a feature limitation.

5. Conclusion

Two features derived from a single PPG sensor — heart rate and inter-beat interval — achieved 68.01 ± 5.82% accuracy on three-class exercise classification, statistically indistinguishable from the full five-feature model (permutation test $p = 0.215$). The pre-registered hypothesis pair HR + EDA also performed within the 5-percentage-point sufficiency threshold ($p = 0.620$). Per-class analysis revealed that the classifier primarily distinguishes resting (stress-induction) sessions from exercise (F1 = 0.79–0.84) rather than performing true three-class discrimination (aerobic and anaerobic F1 = 0.33–0.67). For this rest-versus-exercise detection task, a single PPG sensor providing HR and IBI may be sufficient, and dedicated EDA, temperature, and BVP sensors could potentially be omitted without statistically significant accuracy loss. These findings apply to session-level aggregate features; real-time classification or finer exercise subtype discrimination would likely require temporal feature engineering, accelerometer data, and substantially larger participant samples.

References

1. Fitness Tracker Market Size & Share. https://www.grandviewresearch.com/industry-analysis/fitness-tracker-market.
2. Website. E4 wristband | Real-time physiological signals Empatica https://www.empatica.com › en-eu › research › e4.
3. Shaffer, F. & Ginsberg, J. P. An Overview of Heart Rate Variability Metrics and Norms. *Front Public Health* **5**, 258 (2017).
4. Iqbal, T. *et al.* A Review of Biophysiological and Biochemical Indicators of Stress for Connected and Preventive Healthcare. *Diagnostics (Basel)* **11**, (2021).
5. Healey, J. A. & Picard, R. W. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Trans. Intell. Transp. Syst.* **6**, 156–166 (2005).
6. Banos, O., Galvez, J.-M., Damas, M., Pomares, H. & Rojas, I. Window size impact in human activity recognition. *Sensors (Basel)* **14**, 6474–6499 (2014).
7. Patel, S., Park, H., Bonato, P., Chan, L. & Rodgers, M. A review of wearable sensors and systems with application in rehabilitation. *J Neuroeng Rehabil* **9**, 21 (2012).
8. Hongn, A., Bosch, F., Prado, L. & Bonomini, P. Wearable Device Dataset from Induced Stress and Structured Exercise Sessions. https://doi.org/10.13026/he0v-tf17 (2025).
9. Hongn, A., Bosch, F., Prado, L. E., Ferrández, J. M. & Bonomini, M. P. Wearable Physiological Signals under Acute Stress and Exercise Conditions. *Sci Data* **12**, 520 (2025).
10. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. (2012).