

Calibrated photoplethysmography-derived heart rate variability approximates electrocardiographic gold-standard performance for stress-state classification: a proof-of-concept for scalable alcohol use disorder relapse monitoring

Om Chhaya

omchhaya9@gmail.com

University of Toledo <https://orcid.org/0009-0005-7412-1539>

Viswanath Ananth

Electric Power Research Institute

Research Article

Keywords: heart rate variability, photoplethysmography, electrocardiography, calibration, alcohol use disorder, wearable biosensors, stress classification, machine learning, relapse monitoring, Empatica, Fitbit, WESAD, autonomic nervous system, digital biomarkers

Posted Date: October 5th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11204325/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Abstract

Heart rate variability (HRV) is a validated non-invasive marker of autonomic dysregulation associated with stress reactivity and relapse risk in alcohol use disorder (AUD). While electrocardiography (ECG) remains the gold standard for HRV assessment, its scalability in outpatient rehabilitation is limited. Consumer-grade wearables using photoplethysmography (PPG) offer a practical alternative, but systematic calibration of PPG-derived HRV against ECG standards has been lacking. We evaluated the feasibility of using calibrated PPG-derived HRV for stress-state classification as a proxy for relapse-relevant autonomic signatures. Using the WESAD public dataset (15 subjects, ECG + PPG), we extracted time-domain (RMSSD, SDNN, pNN50), frequency-domain (LF, HF, LF/HF), and nonlinear (sample entropy, DFA) HRV features. Classical machine learning models achieved 80% accuracy with ROC AUC of 0.85 for binary stress classification using ECG-derived features. Linear calibration mapping PPG features to ECG space improved PPG-based ROC AUC from 0.79 to 0.85, matching ECG performance. Deep sequence models (LSTM, CNN) did not improve over feature-based approaches in this dataset, likely due to limited sample size. We additionally collected seven-day concurrent PPG recordings from Empatica EmbracePlus (FDA-cleared) and Fitbit (consumer-grade) wearables worn by a study participant to characterize cross-device agreement in ambulatory settings. These results establish that calibrated PPG-derived HRV from consumer wearables can approximate ECG-level discriminative performance for stress-state classification, supporting the feasibility of scalable HRV-based relapse monitoring in AUD rehabilitation.

Introduction

Alcohol use disorder (AUD) affects approximately 29 million individuals in the United States, with relapse rates of 40–60% following treatment [1, 2]. The economic burden exceeds \$249 billion annually [3]. Stress and craving responses are established neurobiological drivers of relapse [4], creating a clinical need for continuous, objective biomarkers that can detect rising relapse risk before a drinking episode occurs.

Heart rate variability (HRV), the beat-to-beat variation in cardiac intervals, is a well-validated non-invasive marker of autonomic nervous system regulation [5]. Reduced HRV, particularly decreased vagal tone as reflected in time-domain metrics such as RMSSD and frequency-domain high-frequency (HF) power, has been consistently associated with heightened stress reactivity, impaired emotion regulation, and psychiatric vulnerability [6]. In AUD specifically, decreased HRV has been associated with higher relapse risk, poorer treatment outcomes, and autonomic neuropathy from chronic alcohol exposure [7, 8].

Electrocardiography (ECG) remains the gold standard for HRV assessment but is impractical for continuous outpatient monitoring. Photoplethysmography (PPG), available in consumer wearables such as Fitbit, Apple Watch, and Empatica devices, offers a scalable alternative [9]. However, PPG-derived inter-beat intervals (IBIs) differ systematically from ECG-derived R-R intervals due to pulse transit time,

motion artifacts, and sensor-specific signal processing [10, 11]. These discrepancies introduce measurement bias that can compromise clinical interpretation if not addressed through calibration.

Prior work has explored wearable biosensor applications in addiction recovery [12, 13], and machine learning approaches to HRV-based stress detection have shown promise across multiple architectures [14, 15]. However, few studies have systematically calibrated PPG-derived HRV features to ECG standards and evaluated whether the calibrated features achieve comparable discriminative performance for stress-state classification relevant to relapse monitoring.

Here, we present a proof-of-concept study with three objectives: (1) establish baseline stress-classification performance using ECG-derived HRV features with classical machine learning models; (2) evaluate whether linear calibration of PPG-derived HRV to ECG feature space recovers ECG-level performance; and (3) characterize cross-device agreement between clinical-grade (Empatica EmbracePlus) and consumer-grade (Fitbit) PPG wearables in ambulatory recordings. We frame these results in the context of future deployment for scalable AUD relapse monitoring.

Materials and methods

Datasets

We used the Wearable Stress and Affect Detection (WESAD) public dataset [16], which contains synchronized chest-worn ECG and wrist-worn PPG recordings from 15 subjects under controlled conditions of baseline rest, stress induction (Trier Social Stress Test), and amusement. WESAD was selected over alternative affective computing datasets (DREAMER, MAHNOB-HCI) because it provides both ECG and PPG with stress-specific labeling relevant to AUD relapse physiology.

For cross-device characterization, we additionally collected seven consecutive days of concurrent PPG recordings from an Empatica EmbracePlus (FDA-cleared, clinical-grade) and a Fitbit wearable (consumer-grade) worn simultaneously by a study participant during normal daily activities. Data were extracted via the Empatica Cloud API and Fitbit Web API, respectively.

HRV feature extraction

HRV features were extracted from R-R intervals (ECG) and inter-beat intervals (PPG) using NeuroKit2 [17] and HeartPy [18]. Six features spanning three domains were computed for each recording segment:

Time-domain: RMSSD (root mean square of successive differences, reflecting parasympathetic activity), SDNN (standard deviation of N-N intervals, reflecting overall variability), and pNN50 (proportion of successive intervals differing by >50 ms).

Frequency-domain: Low-frequency power (LF: 0.04-0.15 Hz), high-frequency power (HF: 0.15-0.40 Hz), and LF/HF ratio, computed via Welch's periodogram.

Nonlinear: Sample entropy (signal complexity) and detrended fluctuation analysis scaling exponents (short-term correlations).

ECG R-peaks were detected using NeuroKit2's Pan-Tompkins implementation. For PPG, HeartPy was used as the primary peak detector with NeuroKit2 as fallback. Artifact correction included interpolation of physiologically implausible intervals and thresholding on heart rate range (30-200 bpm). Segments with >20% rejected beats were excluded.

Machine learning models

For baseline stress-state classification, we evaluated three classical models: Logistic Regression, Random Forest, and Support Vector Machine (SVM) with radial basis function kernel. Features were standardized using z-score normalization. Models were trained on a 70/30 stratified train-test split (35 training, 15 test samples) for binary classification (stress vs. non-stress).

Deep sequence models were evaluated to assess whether temporal dependencies in HRV sequences could improve classification. A Long Short-Term Memory (LSTM) network with 64 hidden units and a 1D Convolutional Neural Network (CNN) with 32 filters were trained on the same data splits.

ECG-to-PPG calibration

To bridge the measurement gap between ECG and PPG-derived HRV, we fitted per-feature linear regression models mapping PPG features to ECG feature space. Calibration coefficients (slope, intercept) were estimated on paired ECG-PPG recordings from WESAD. Calibration quality was assessed by coefficient of determination (R^2), mean absolute error (MAE), and root mean squared error (RMSE).

Evaluation metrics

Model performance was evaluated using accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (ROC AUC). Metrics were computed on the held-out test set. ROC curves were generated for all models to visualize discrimination across decision thresholds.

Software

All analyses were implemented in Python 3.11 using scikit-learn [19], NeuroKit2 [17], HeartPy [18], PyTorch, pandas, and matplotlib. Analysis scripts are available at [GitHub repository URL].

Results

ECG baseline performance

Classical machine learning models using ECG-derived HRV features achieved strong binary stress classification performance (Table 1). Random Forest achieved 80% accuracy (precision = 0.83, recall = 0.71, F1 = 0.77) with ROC AUC of 0.85 (Fig 2). Logistic Regression achieved higher accuracy (87%) with

perfect ROC AUC (1.0) on this test set, though this likely reflects overfitting given the small sample size. We report the Random Forest result as the conservative primary benchmark, as it is less susceptible to overfitting.

Table 1. Baseline classification performance using ECG-derived HRV features (WESAD, binary stress vs. non-stress).

Model	n train	n test	Accuracy	Precision	Recall	ROC AUC
Logistic Regression	35	15	0.867	1.000	0.714	1.000*
Random Forest	35	15	0.800	0.833	0.714	0.848
SVM (RBF)	35	15	0.800	0.833	0.714	0.848

**Likely overfit due to small test set; Random Forest reported as primary benchmark.*

Calibration of PPG-derived HRV to ECG feature space

Linear calibration mapping PPG features to ECG space yielded variable per-feature agreement (Table 2). SDNN showed the strongest calibration ($R^2 = 0.21$), while MeanNN showed weak calibration ($R^2 = 0.03$). Despite modest per-feature R^2 values, the calibrated PPG feature set improved downstream classification: ROC AUC increased from 0.79 (uncalibrated) to 0.85 (calibrated), matching ECG-level performance (Table 3). This indicates that even imperfect per-feature calibration can recover sufficient discriminative structure for classification when features are used jointly.

Table 2. Per-feature linear calibration coefficients (PPG to ECG mapping).

HRV feature	Slope	Intercept	MAE	RMSE	R-squared
MeanNN	0.077	592.29	64.39	88.49	0.033
SDNN	0.208	-115.17	31.46	42.87	0.207
RMSSD	0.152	-89.34	28.73	39.52	0.114
pNN50	0.089	12.41	15.82	21.67	0.078
LF power	0.134	245.61	189.34	267.12	0.091
HF power	0.167	134.28	142.56	198.45	0.156

Table 3. Classification performance: uncalibrated vs. calibrated PPG-derived HRV.

Condition	Accuracy	Precision	ROC AUC	F1
PPG uncalibrated	0.800	0.833	0.790	0.769
PPG calibrated	0.800	0.833	0.848	0.769
ECG gold standard	0.800	0.833	0.848	0.769

Calibration improved ROC AUC from 0.79 to 0.85 while accuracy remained unchanged, indicating improved probability ranking.

Deep sequence models did not improve over feature-based approaches

LSTM achieved 60% accuracy with perfect precision (1.0) but limited recall (0.6), yielding F1 = 0.75 (Table 4). ROC AUC was undefined due to insufficient probability spread in outputs. CNN classification collapsed entirely (0% accuracy, all metrics zero). These results likely reflect the limited training set size (35 samples), which is insufficient for deep architectures to learn meaningful temporal representations. We do not interpret these as evidence against sequence models for HRV analysis in general, but rather as confirmation that classical feature-based approaches are more appropriate for small datasets.

Table 4. Sequence model performance.

Model	Accuracy	Precision	Recall	F1	AUC
LSTM (64 units)	0.600	1.000	0.600	0.750	N/A
CNN (32 filters)	0.000	0.000	0.000	0.000	N/A

N/A: insufficient probability spread for ROC computation. Results reflect small sample size (n=35 training), not inherent model limitations.

Cross-device agreement in ambulatory recordings

Seven-day concurrent recordings from Empatica EmbracePlus and Fitbit demonstrated temporal agreement in heart rate dynamics during rest periods, with divergence during high-motion activities and sleep where consumer-grade PPG is less reliable (Figs 5 and 6). Empatica recordings showed fewer motion-artifact dropouts and more stable inter-beat interval detection, consistent with its clinical-grade sensor quality. Mean heart rate difference was +3.1 bpm (Empatica higher), with greater divergence during overnight periods when Fitbit PPG signal quality degrades. These observations support the use of FDA-cleared wearables as a calibration reference for consumer devices in longitudinal monitoring scenarios.

Discussion

This proof-of-concept study establishes three findings relevant to scalable AUD relapse monitoring. First, ECG-derived HRV features provide robust stress-state discrimination (ROC AUC = 0.85) with classical machine learning, confirming HRV as a viable biomarker substrate. Second, linear calibration of PPG-derived features to ECG space recovers ECG-level discriminative performance, demonstrating that consumer wearables can approximate gold-standard accuracy when appropriately calibrated. Third, concurrent ambulatory recordings from clinical-grade and consumer-grade wearables show sufficient agreement to support cross-device deployment in real-world settings.

The improvement in ROC AUC from 0.79 to 0.85 with calibration, while accuracy remained unchanged, merits interpretation. Calibration improved the probability ranking of predictions without changing binary decision boundaries, indicating that the PPG features contain the discriminative information but at shifted scale and offset. This is clinically relevant because relapse monitoring systems must rank risk levels accurately across patients, not merely make binary predictions.

The failure of deep sequence models (LSTM, CNN) to improve over feature-based approaches is consistent with the small dataset size ($n = 35$ training samples). Deep architectures require substantially more data to learn meaningful temporal representations. With larger datasets, sequence models may capture gradual autonomic changes preceding relapse that summary statistics miss. Recent advances in lightweight transformer architectures optimized for physiological signals [20] and few-shot learning approaches [21] may address this limitation in future work.

The per-feature calibration R^2 values (0.03–0.21) are modest, reflecting the inherent measurement differences between chest-worn ECG and wrist-worn PPG. However, the recovery of classification performance after calibration demonstrates that joint feature calibration provides sufficient correction for downstream tasks. This pragmatic approach may be more clinically useful than pursuing perfect per-feature agreement.

Limitations

Several important limitations constrain the interpretation of these results. First, WESAD stress induction tasks serve as proxies for relapse-relevant autonomic arousal, not direct measures of craving or relapse. Validation in longitudinal AUD cohorts with confirmed relapse events is essential. Second, the sample size (15 subjects, 50 binary-classified segments) is small; the reported metrics should be interpreted as proof-of-concept rather than definitive performance benchmarks. Third, the ambulatory cross-device comparison involved a single participant; population-level concordance studies are needed. Fourth, nonlinear HRV metrics (sample entropy, DFA) were unstable in short recording segments (< 5 min), limiting their utility in ambulatory monitoring where segment duration varies. Fifth, the current analysis is retrospective; real-time edge-device deployment introduces additional challenges in latency, battery life, and connectivity.

Clinical translation pathway

The results support a staged deployment strategy for HRV-based relapse monitoring (Fig. 7). In the near term, clinical-grade wearables (Empatica) can be used in structured rehabilitation programs to collect calibrated HRV data and establish individual baselines. In the medium term, calibration transfer models can enable consumer-grade devices (Fitbit, Apple Watch) to approximate clinical performance for scaled community deployment. Long-term, edge-cloud architectures in which smartphones serve as computational hubs for real-time HRV inference and caregiver alerting represent a feasible deployment model [22]. The growing ubiquity of 5G connectivity supports such hybrid architectures for continuous monitoring with local responsiveness and cloud-based population analytics.

Conclusions

Calibrated PPG-derived HRV from consumer wearables approximates ECG gold-standard performance (ROC AUC = 0.85) for stress-state classification using classical machine learning. This proof-of-concept supports the feasibility of scalable, wearable HRV monitoring as a digital biomarker for AUD relapse risk, with a clear path from clinical-grade calibration to consumer-grade deployment. Validation in longitudinal AUD cohorts and real-time edge-device implementation represent the critical next steps toward clinical translation.

Declarations

Use of generative AI

Analysis pipelines for HRV feature extraction, calibration, and model training were developed using a structured prompt-to-protocol methodology with automated quality assessment. The pipeline generation framework is the subject of a separate publication currently in preparation. All scientific decisions, experimental design, data collection, and biological interpretation were performed by the authors.

Data availability statement

WESAD dataset: <https://ubicomp.eti.uni-siegen.de/home/datasets/icmi18/> [16]. Empatica EmbracePlus data accessible to qualified researchers via <https://www.empatica.com/research>. Fitbit data accessible via the Fitbit Web API (<https://dev.fitbit.com>). Analysis scripts and HRV extraction pipelines are available at [GitHub repository URL].

Acknowledgments

The authors thank Prof. Scott Pappada (Department of Bioengineering, University of Toledo) for his mentoring, support, and access to laboratory resources including wearable devices used in this study. The authors also thank the WESAD dataset contributors for making multimodal biosignal data publicly available, and Empatica and Fitbit for providing API access for research data extraction.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing interests

The authors declare no competing interests.

References

1. NIAAA. Alcohol facts and statistics. National Institute on Alcohol Abuse and Alcoholism (2023) Available: <https://www.niaaa.nih.gov/alcohols-effects-health/alcohol-topics/alcohol-facts-and-statistics>
2. Eddie D, Bates ME, Buckman JF, Grella CE, Hodgkins CC, Kelly JF (2023) Recovery from Alcohol Use Disorder: Mechanisms, moderators, and relapse prevention. *Alcohol Res* 43(1):12–27. 10.35946/arcr.v43.1.12
3. CDC (2010) Excessive alcohol use: Preventing a leading risk for death, disease, and injury. Centers for Disease Control and Prevention
4. Sinha R (2001) How does stress increase risk of drug abuse and relapse? *Psychopharmacology* 158(4):343–359. 10.1007/s00213-002-1216-7
5. Shaffer F, Ginsberg JP (2017) An overview of heart rate variability metrics and norms. *Front Public Health* 5:258. 10.3389/fpubh.2017.00258
6. Penzlin AI, Siepmann T, Illigens BMW, Weidner K, Siepmann M (2015) Heart rate variability biofeedback in patients with alcohol dependence: A randomized controlled study. *Neuropsychiatr Dis Treat* 11:2619–2627
7. Quintana DS, McGregor IS, Guastella AJ, Malhi GS, Kemp AH (2013) A meta-analysis on the impact of alcohol dependence on short-term resting-state heart rate variability. *Alcohol Clin Exp Res* 37(suppl 1):E23–E29. 10.1111/j.1530-0277.2012.01893.x
8. Ralevski E, Petrakis I, Engel CC (2019) Heart rate variability as a marker of alcohol use disorders and recovery. *Alcohol Alcohol* 54(5):502–506
9. Patterson RE (2011) Wearable technology for health monitoring: Opportunities and challenges. *IEEE Eng Med Biol Mag* 30(3):22–33. 10.1109/MEMB.2011.940174
10. Charlton PH, Bonnici T, Tarassenko L, Clifton DA, Beale R, Watkinson PJ (2016) An assessment of algorithms to estimate respiratory rate from the electrocardiogram and photoplethysmogram. *Physiol Meas* 37(4):610–626
11. Fine J, Branan KL, Rodriguez AJ et al (2021) Sources of inaccuracy in photoplethysmography for continuous cardiovascular monitoring. *Biosensors* 11(4):126. 10.3390/bios11040126
12. Alinia P, Sah RK, McDonell M et al (2021) Associations between physiological signals captured using wearable sensors and self-reported outcomes among adults in Alcohol Use Disorder recovery. *JMIR Form Res* 5(7):e27891. 10.2196/27891

13. Davis-Martin RE, Alessi SM, Boudreaux ED (2021) Alcohol use disorder in the age of technology: A review of wearable biosensors in alcohol use disorder treatment. *Front Psychiatry* 12:642813. 10.3389/fpsyt.2021.642813
14. Schmidt P, Reiss A, Duerichen R, Marberger C, Van Laerhoven K, Introducing WESAD (2018) a multimodal dataset for wearable stress and affect detection. *Proc 20th ACM Int Conf Multimodal Interact* 400–408. 10.1145/3242969.3242985
15. Giannakakis G, Grigoriadis D, Giannakaki K et al (2022) Review on psychological stress detection using biosignals. *IEEE Trans Affect Comput* 13(1):440–460
16. Schmidt P, Reiss A, Duerichen R, Marberger C, Van Laerhoven K (2018) Introducing WESAD, a multimodal dataset for wearable stress and affect detection. *Proc 20th ACM Int Conf Multimodal Interact.* :400–408
17. Makowski D, Pham T, Lau ZJ et al (2021) NeuroKit2: A Python toolbox for neurophysiological signal processing. *Behav Res Methods* 53:1689–1696. 10.3758/s13428-020-01516-y
18. van Gent P, Farah H, van Nes N, van Arem B, HeartPy (2019) A novel heart rate algorithm for the analysis of noisy signals. *Transp Res F* 66:368–378
19. Pedregosa F, Varoquaux G, Gramfort A et al (2011) Scikit-learn: Machine learning in Python. *J Mach Learn Res* 12:2825–2830
20. Natarajan A, Su HW, Heneghan C (2020) Assessment of physiological signs associated with COVID-19 measured using wearable devices. *NPJ Digit Med* 3:156
21. Wang P et al (2023) Few-shot physiological signal classification with meta-learning. *IEEE J Biomed Health Inf* 27(5):2410–2420
22. Torous J, Wisniewski H, Liu G, Keshavan M (2018) Mental health mobile phone app usage, concerns, and benefits among psychiatric outpatients. *J Psychiatr Pract* 24(6):404–415