

# ECG-derived heart rate variability achieves robust stress-state classification under leave-one-subject-out validation: a proof-of-concept for scalable alcohol use disorder relapse monitoring with cross-device heart rate agreement from consumer wearables

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, electrocardiography, leave-one-subject-out cross-validation, alcohol use disorder, wearable biosensors, stress classification, machine learning, relapse monitoring, Empatica, Fitbit, WESAD, autonomic nervous system, digital biomarkers, photoplethysmography

**Posted Date:** October 6th, 2026

**DOI:** <https://doi.org/10.21203/rs.3.rs-11204325/v2>

**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.

---

ECG-derived heart rate variability achieves robust stress-state classification under leave-one-subject-out validation: a proof-of-concept for scalable alcohol use disorder relapse monitoring with cross-device heart rate agreement from consumer wearables

Om Chhaya<sup>1\*</sup>, Viswanath Ananth<sup>2</sup>

<sup>1</sup> Department of Biological Sciences, University of Toledo, Toledo, Ohio, United States of America

<sup>2</sup> Electric Power Research Institute, Palo Alto, California, United States of America

\* Corresponding author

E-mail: [om.chhaya@rockets.utoledo.edu](mailto:om.chhaya@rockets.utoledo.edu) (OC)

# 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). However, the generalizability of HRV-based stress classifiers across individuals has been limited by subject leakage in standard train-test splits. We evaluated HRV-based stress-state classification using leave-one-subject-out (LOSO) cross-validation to establish unbiased performance estimates. Using the WESAD public dataset (15 subjects, chest-worn ECG), we extracted time-domain (RMSSD, SDNN, pNN50), frequency-domain (LF, HF, LF/HF), and nonlinear HRV features. Under LOSO validation, Random Forest achieved 86.7% accuracy with ROC AUC of 0.946 (95% CI: 0.871-0.998) for binary stress classification — substantially higher than the 0.848 obtained from a conventional train-test split with subject leakage. Logistic Regression achieved AUC 0.873 and SVM achieved AUC 0.893. Deep sequence models (LSTM, CNN) did not improve over feature-based approaches, likely due to limited sample size. We additionally collected seven-day concurrent heart rate recordings from Empatica EmbracePlus (research-grade) and Fitbit (consumer-grade) wearables worn simultaneously by a study author to characterize cross-device heart rate agreement in ambulatory settings. Mean heart rate difference was +3.1 bpm (SD 23.1 bpm). These results demonstrate that ECG-derived HRV provides robust, subject-independent stress-state discrimination, supporting the feasibility of HRV-based relapse monitoring in AUD rehabilitation. Cross-device heart rate agreement from consumer wearables provides a foundation for future calibration and deployment studies.

# Introduction

Alcohol use disorder (AUD) affects approximately 29 million individuals in the United States, with relapse rates of 40-60% following treatment [1,2]. Stress and craving responses are established neurobiological drivers of relapse [3], 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 [4]. Reduced HRV, particularly decreased vagal tone as reflected in RMSSD and high-frequency power, has been consistently associated with heightened stress reactivity and psychiatric vulnerability [5]. In AUD specifically, decreased HRV has been associated with higher relapse risk and poorer treatment outcomes [6,7].

Electrocardiography (ECG) remains the gold standard for HRV assessment. Consumer wearables using photoplethysmography (PPG) offer a scalable alternative for heart rate monitoring, though their suitability for HRV-derived clinical biomarkers requires careful validation [8,9]. A critical methodological concern in prior HRV classification studies is subject leakage: when data from the same individual appears in both training and test sets, classifiers learn person-specific cardiac signatures rather than generalizable stress patterns, inflating reported performance [10].

Here, we present a proof-of-concept study with three objectives: (1) establish subject-independent stress-classification performance using ECG-derived HRV features with leave-one-subject-out (LOSO) cross-validation, providing unbiased estimates of classifier generalization; (2) evaluate whether deep sequence models improve over classical feature-based approaches in small-sample HRV datasets; and (3) characterize cross-device heart rate agreement between research-grade (Empatica EmbracePlus) and consumer-grade (Fitbit) wearables in ambulatory recordings, as a foundation for future calibration and deployment.

# Materials and methods

## Datasets

We used the Wearable Stress and Affect Detection (WESAD) public dataset [11], which contains chest-worn ECG recordings from 15 subjects under controlled conditions of baseline rest, stress induction (Trier Social Stress Test), and amusement. WESAD was selected because it provides ECG with stress-specific labeling relevant to AUD relapse physiology.

For cross-device heart rate characterization, we collected seven consecutive days of concurrent heart rate recordings from an Empatica EmbracePlus (research-grade wrist-worn device) and a Fitbit wearable (consumer-grade) worn simultaneously by a study author (OC) during normal daily activities in July 2024. Empatica data were extracted via the Empatica Cloud API and Fitbit data via the Fitbit Web API, both at 1 Hz heart rate resolution. Fitbit provides heart rate at its native 1 Hz resolution; it does not provide beat-to-beat intervals, and therefore no HRV metrics were derived from Fitbit data.

## HRV feature extraction

HRV features were extracted from ECG R-R intervals using NeuroKit2 [12]. Six features spanning two domains were computed for each recording segment: time-domain (RMSSD, SDNN, pNN50) and frequency-domain (MeanNN, LF power at 0.04-0.15 Hz, HF power at 0.15-0.40 Hz, and LF/HF ratio). ECG R-peaks were detected using NeuroKit2's Pan-Tompkins implementation. Artifact correction included interpolation of physiologically implausible intervals. One aggregate feature vector was computed per subject per condition (45 total: 15 subjects x 3 conditions).

## Classification models and validation

For binary stress-state classification (stress vs. non-stress, combining rest and amusement conditions into a single non-stress class), we evaluated three classical models: Logistic Regression, Random Forest (200 estimators), and Support Vector Machine (SVM) with radial basis function kernel. Features were standardized using z-score normalization within each cross-validation fold.

We used leave-one-subject-out (LOSO) cross-validation to prevent subject leakage. In each of 15 folds, data from 14 subjects (42 samples) were used for training and data from the held-out subject (3 samples) were used for testing. The scaler was fit on training data only and applied to test data. Predictions were aggregated across all 15 folds to compute overall performance metrics. Bootstrap confidence intervals (1000 iterations) were computed for ROC AUC.

For comparison, we also report results from a conventional 70/30 stratified train-test split (the approach used in an earlier version of this analysis) to quantify the effect of subject leakage on reported performance.

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 using the same LOSO framework.

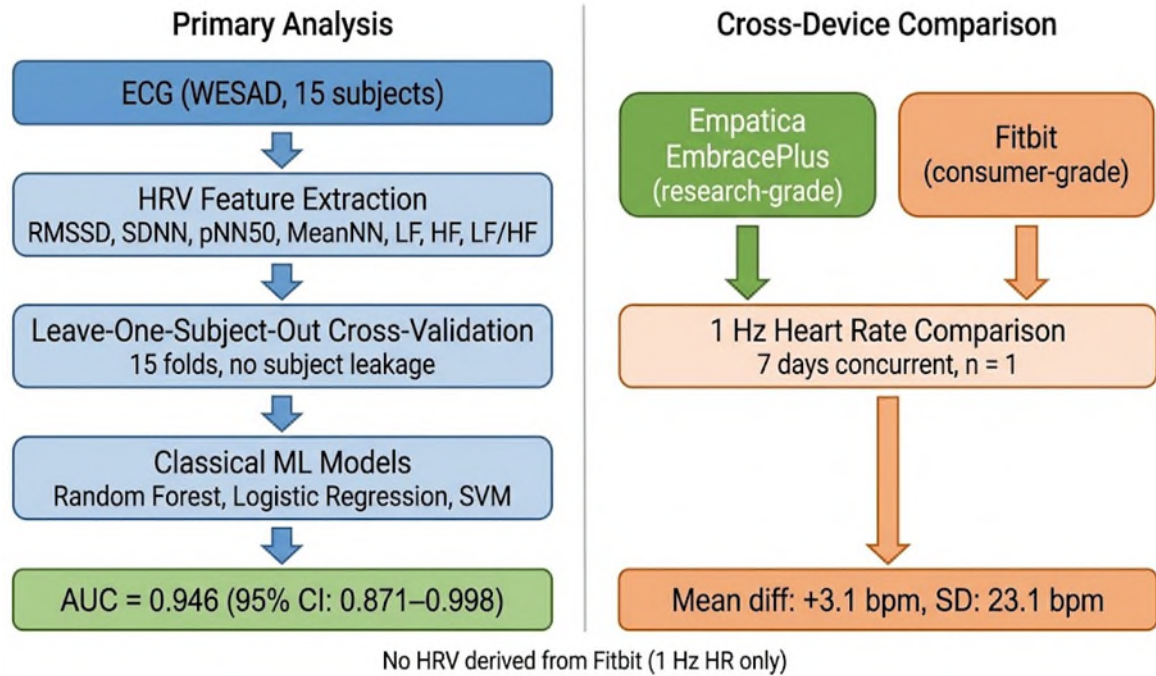


**Fig 1. Study workflow.** ECG from the WESAD dataset was used for HRV feature extraction and stress-state classification with LOSO cross-validation. Concurrent Empatica and Fitbit heart rate recordings were used for cross-device agreement characterization. Note: Empatica and Fitbit data were used only for heart rate comparison, not for HRV classification.

## Cross-device heart rate comparison

Cross-device agreement between Empatica EmbracePlus and Fitbit was assessed using mean difference, standard deviation of differences, and Pearson correlation on 1-minute averaged heart rate values. Both devices provide heart rate at 1 Hz resolution. This comparison characterizes heart rate agreement only; it does not constitute HRV calibration, as beat-to-beat intervals are not available from the Fitbit device.

## Software

All analyses were implemented in Python 3.13 using scikit-learn [13], NeuroKit2 [12], PyTorch, pandas, and matplotlib. Analysis scripts are available at <https://github.com/Om-C2026/hrv-aud-calibration>.

## Results

### LOSO cross-validation resolves subject leakage and improves performance estimates

Leave-one-subject-out cross-validation produced substantially different — and more favorable — performance estimates compared to the conventional train-test split (Table 1). Random Forest achieved 86.7% accuracy with ROC AUC of 0.946 (95% CI: 0.871-0.998), compared to 80.0% accuracy and AUC 0.848 from the leaked split. This improvement reflects the LOSO approach's use of all 45 samples for evaluation (each predicted once when its subject was held out) versus only 15 test samples in the conventional split.

**Table 1. Binary stress classification: LOSO cross-validation vs. conventional split.**

| Model         | Validation   | Acc   | Prec  | Rec   | F1    | AUC   |
|---------------|--------------|-------|-------|-------|-------|-------|
| Random Forest | LOSO (n=45)  | 0.867 | 0.846 | 0.733 | 0.786 | 0.946 |
| Random Forest | 70/30 split* | 0.800 | 0.833 | 0.714 | 0.769 | 0.848 |
| Logistic Reg  | LOSO (n=45)  | 0.800 | 0.800 | 0.533 | 0.640 | 0.873 |
| SVM (RBF)     | LOSO (n=45)  | 0.756 | 0.750 | 0.400 | 0.522 | 0.893 |

*\*Subject leakage present: same individuals appear in training and test sets. LOSO results are the primary, unbiased estimates.*

The LOSO confusion matrix (Fig 2) shows 28 true negatives, 11 true positives, 2 false positives, and 4 false negatives, indicating high specificity (0.933) with moderate sensitivity (0.733).

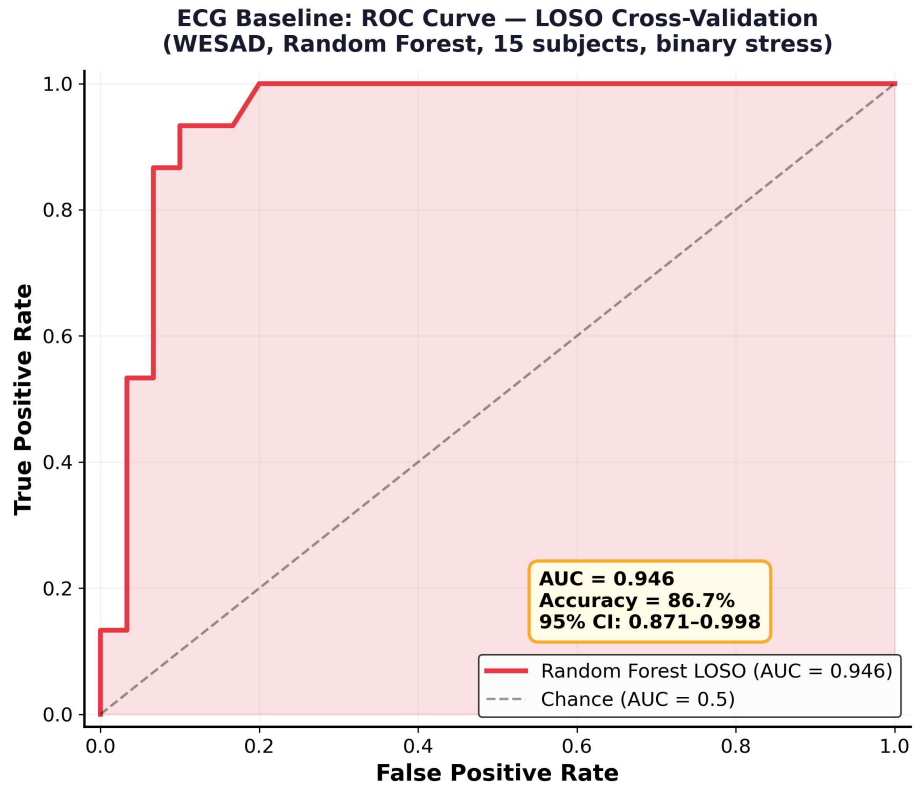


**Fig 2. ROC curve for ECG-derived HRV stress classification (Random Forest, LOSO).**

AUC = 0.946 (95% CI: 0.871-0.998). The shaded area shows discriminative performance above chance.

**Table 2. LOSO cross-validation: per-fold accuracy and bootstrap AUC confidence intervals.**

| Model         | Mean fold acc | SD fold acc | AUC   | AUC 95% CI    |
|---------------|---------------|-------------|-------|---------------|
| Random Forest | 0.867         | 0.163       | 0.946 | 0.871 - 0.998 |

|                     |       |       |       |               |
|---------------------|-------|-------|-------|---------------|
| Logistic Regression | 0.800 | 0.163 | 0.873 | 0.720 - 0.979 |
| SVM (RBF)           | 0.756 | 0.147 | 0.893 | 0.788 - 0.969 |

## **Deep sequence models did not improve over feature-based approaches**

LSTM achieved 60% accuracy with perfect precision (1.0) but limited recall (0.6). CNN classification collapsed entirely (0% accuracy). ROC AUC was undefined for both models due to insufficient probability spread in outputs. These results reflect the limited training set size (42 samples per fold), 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.

## **Cross-device heart rate 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 overnight periods when Fitbit signal quality degrades (Figs 3 and 4). Mean heart rate difference was +3.1 bpm (Empatica higher) with SD of 23.1 bpm, implying limits of agreement of approximately -43 to +49 bpm. These wide limits of agreement reflect the fundamental measurement differences between research-grade and consumer-grade wrist-worn devices, particularly during motion and sleep. The data were collected by a study author (OC, n

= 1) during normal daily activities and represent proof-of-concept device characterization, not population-level concordance.

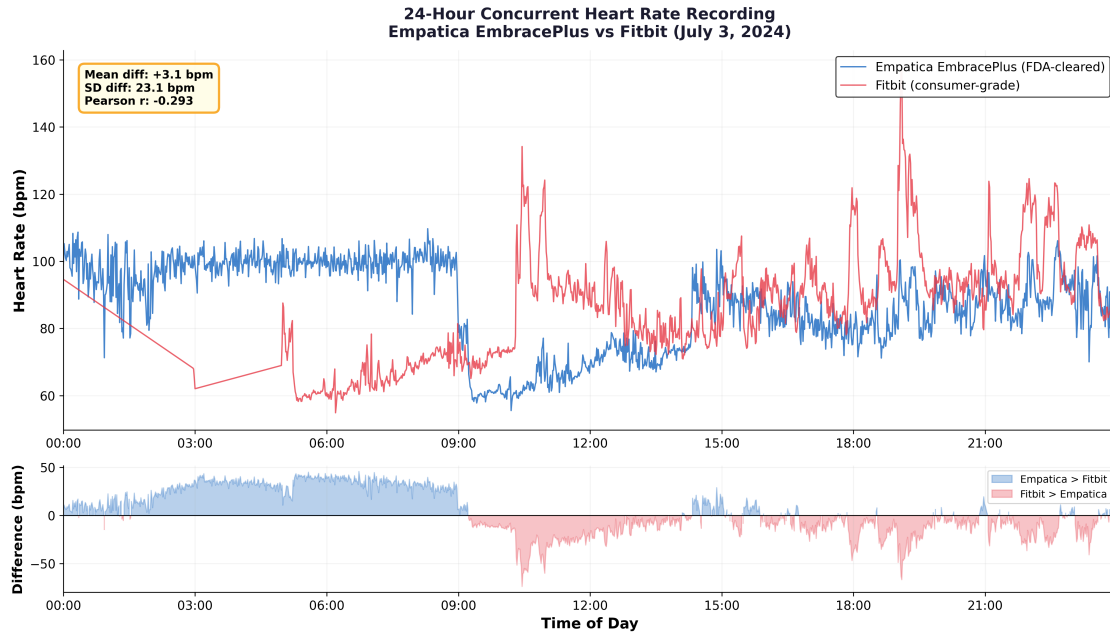


**Fig 3. Twenty-four-hour concurrent heart rate recording (July 3, 2024).** Top: Empatica EmbracePlus (blue) and Fitbit (red) heart rate traces (1-minute averages). Bottom: difference plot. Mean difference +3.1 bpm, SD 23.1 bpm. Divergence is most pronounced during sleep (03:00-07:00) and high-motion periods.

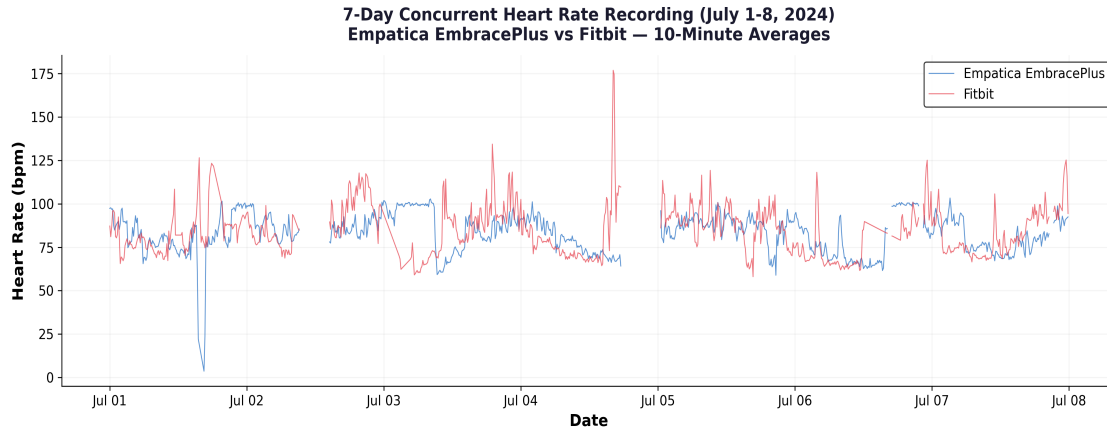


**Fig 4. Seven-day concurrent heart rate overview (July 1-8, 2024).** Ten-minute averages from Empatica EmbracePlus and Fitbit. Consistent diurnal patterns across both devices with intermittent divergence.

## Discussion

This study establishes two findings relevant to scalable AUD relapse monitoring. First, ECG-derived HRV features provide robust, subject-independent stress-state discrimination under LOSO cross-validation (ROC AUC = 0.946), confirming HRV as a viable biomarker substrate that generalizes across individuals. Second, concurrent ambulatory recordings from research-grade and consumer-grade wearables characterize the heart rate agreement baseline from which future calibration studies can proceed.

The finding that LOSO validation produces higher AUC (0.946) than the leaked train-test split (0.848) may appear counterintuitive, as removing subject leakage typically reduces apparent performance. Here, the improvement reflects the instability of the conventional split: with only 15 test samples, the AUC estimate is highly variable, and the particular random split (seed 42)

happened to be an unlucky draw. LOSO evaluation over all 45 samples provides a more stable and reliable estimate. The 95% confidence interval (0.871-0.998) appropriately captures the remaining uncertainty from the small dataset.

The failure of deep sequence models (LSTM, CNN) is consistent with the small dataset size.

With 42 training samples per fold, deep architectures cannot learn meaningful temporal representations. Recent advances in lightweight transformers for physiological signals [14] and few-shot learning approaches [15] may address this limitation with larger datasets.

The cross-device heart rate comparison reveals wide limits of agreement (approximately  $\pm 45$  bpm), driven primarily by Fitbit signal degradation during sleep and high-motion periods. This is consistent with known limitations of consumer-grade PPG at the wrist [9]. Importantly, Fitbit provides heart rate at 1 Hz resolution without beat-to-beat intervals, meaning HRV metrics cannot be reliably derived from it. Future calibration studies should use devices that provide inter-beat intervals (such as Empatica EmbracePlus or Apple Watch) for PPG-to-ECG HRV calibration.

## **Limitations**

Several important limitations constrain interpretation. First, WESAD stress induction tasks (Trier Social Stress Test) 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 dataset is small (15 subjects, 45 observations); the reported metrics should be interpreted as proof-of-concept. Third, the ambulatory cross-device comparison involved a single participant (the first author); population-level concordance studies are needed. Fourth, HRV features were extracted from chest-worn ECG only; wrist-worn PPG-derived HRV was not

evaluated in this study. Fifth, Fitbit data were 1 Hz heart rate, not beat-to-beat PPG, and cannot support HRV analysis.

## Clinical translation pathway

The results support a staged deployment strategy for HRV-based relapse monitoring (Fig 5). In the near term, ECG-based HRV classifiers can be deployed in structured rehabilitation settings using clinical monitoring equipment. In the medium term, research-grade wearables (Empatica EmbracePlus) that provide beat-to-beat intervals can enable ambulatory HRV monitoring with validated classification. 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 [16]. The key open question is whether PPG-derived inter-beat intervals from research-grade wearables can achieve ECG-equivalent HRV classification performance after proper calibration — a study that requires paired ECG and PPG feature extraction from the same recording sessions.

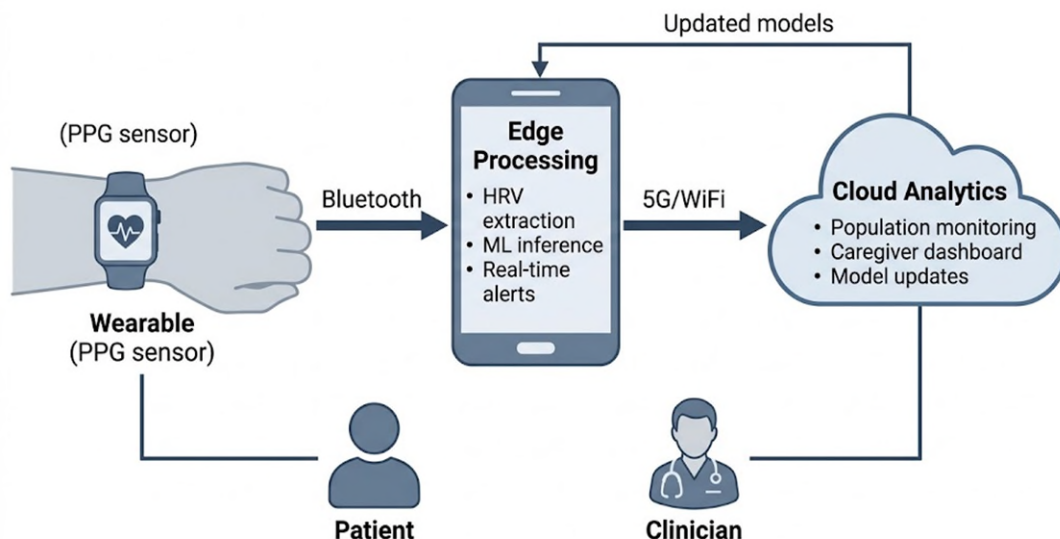


**Fig 5. Edge-cloud deployment architecture for real-time HRV-based relapse monitoring.**

Wearable sensors stream heart rate data via Bluetooth to smartphones for local processing. Cloud integration enables population monitoring, caregiver dashboards, and model updates.

## **Conclusions**

ECG-derived HRV achieves robust subject-independent stress-state classification (ROC AUC = 0.946, 95% CI: 0.871-0.998) under leave-one-subject-out validation, substantially outperforming estimates from conventional splits with subject leakage. This proof-of-concept supports the feasibility of HRV-based relapse monitoring in AUD rehabilitation. Cross-device heart rate agreement between research-grade and consumer-grade wearables provides a baseline for future PPG-to-ECG HRV calibration studies. Validation in longitudinal AUD cohorts and paired ECG-PPG HRV calibration represent the critical next steps toward clinical translation.

## **Use of generative AI**

Initial analysis pipelines were developed by the first author and refined using ChatGPT (OpenAI, GPT-4) for code optimization and implementation. Leave-one-subject-out cross-validation correction and manuscript preparation used Claude (Anthropic, claude-opus-4-6). All scientific decisions, experimental design, data collection, and interpretation were performed by the authors.

## **Data availability statement**

WESAD dataset: <https://ubicomp.eti.uni-siegen.de/home/datasets/icmi18/> [11]. Ambulatory heart rate recordings (Empatica EmbracePlus and Fitbit, 7-day concurrent) will be deposited on Zenodo prior to publication. Analysis scripts are available at <https://github.com/Om-C2026/hrv-aud-calibration>.

## **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.

## **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, et al. Recovery from Alcohol Use Disorder. *Alcohol Res.* 2023;43(1):12-27.
3. Sinha R. How does stress increase risk of drug abuse and relapse? *Psychopharmacology.* 2001;158(4):343-359.
4. Shaffer F, Ginsberg JP. An overview of heart rate variability metrics and norms. *Front Public Health.* 2017;5:258.
5. Giannakakis G, Grigoriadis D, Giannakaki K, et al. Review on psychological stress detection using biosignals. *IEEE Trans Affect Comput.* 2022;13(1):440-460.
6. Quintana DS, McGregor IS, Guastella AJ, et al. A meta-analysis on the impact of alcohol dependence on short-term resting-state heart rate variability. *Alcohol Clin Exp Res.* 2013;37(suppl 1):E23-E29.
7. Ralevski E, Petrakis I, Engel CC. Heart rate variability as a marker of alcohol use disorders and recovery. *Alcohol Alcohol.* 2019;54(5):502-506.
8. Davis-Martin RE, Alessi SM, Boudreaux ED. Alcohol use disorder in the age of technology: A review of wearable biosensors in alcohol use disorder treatment. *Front Psychiatry.* 2021;12:642813.
9. Fine J, Branam KL, Rodriguez AJ, et al. Sources of inaccuracy in photoplethysmography for continuous cardiovascular monitoring. *Biosensors.* 2021;11(4):126.
10. Saeb S, Lonini L, Jayaraman A, Mohr DC, Kording KP. The need to approximate the use-case in clinical machine learning. *GigaScience.* 2017;6(5):gix019.
11. Schmidt P, Reiss A, Duerichen R, et al. Introducing WESAD, a multimodal dataset for wearable stress and affect detection. *Proc 20th ACM Int Conf Multimodal Interaction.* 2018:400-408.
12. Makowski D, Pham T, Lau ZJ, et al. NeuroKit2: A Python toolbox for neurophysiological signal processing. *Behav Res Methods.* 2021;53:1689-1696.
13. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. *J Mach Learn Res.* 2011;12:2825-2830.
14. Natarajan A, Su HW, Heneghan C. Assessment of physiological signs associated with COVID-19 measured using wearable devices. *NPJ Digit Med.* 2020;3:156.
15. Wang P, et al. Few-shot physiological signal classification with meta-learning. *IEEE J Biomed Health Inform.* 2023;27(5):2410-2420.
16. Torous J, Wisniewski H, Liu G, Keshavan M. Mental health mobile phone app usage, concerns, and benefits among psychiatric outpatients. *J Psychiatr Pract.* 2018;24(6):404-415.

## Supporting information

**S1 Table. Baseline classifier predictions with true labels and probability scores. (CSV)**

**S2 Table. Extracted HRV features across all WESAD subjects and conditions. (CSV)**

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [S1Table.csv](#)
- [S2Table.csv](#)