Participant-Independent Identification of Self-Reported Stress Events From Wearable Sensor Data During Alcohol Use Disorder Recovery

A Secondary Observational Machine-Learning Analysis of the ADARP Dataset

Imran Sohail Guller

BASS Emerging Scholars Program

Capstone Research Paper

August 14, 2026

Abstract

Background: Stress is associated with craving and relapse risk during alcohol use disorder (AUD) recovery, but intermittent self-report cannot provide continuous information about stress-related states.

Objective: To evaluate whether changes from an individual's recent physiological baseline could distinguish periods before self-reported stress events from matched control periods in participants excluded from model training, and to compare monitoring windows and sensor modalities.

Design and data source: Secondary analysis of the publicly available Alcohol and Drug Abuse Research Program (ADARP) wearable-sensor dataset.

Participants: Eleven adults receiving treatment for mental health conditions and moderate-to-severe AUD in Washington State.

Predictors and outcome: Twenty-four change-from-baseline features were derived from heart rate, electrodermal activity, skin temperature, and acceleration in 5-, 15-, 30-, and 60-minute windows. The binary outcome was a participant-generated E4 stress marker versus a participant-matched control period.

Analysis: Logistic regression and random forest classifiers were evaluated using leave-one-participant-out validation. Performance was assessed using ROC AUC, average precision, accuracy, F1 score, participant-bootstrap confidence intervals, paired window comparisons, modality ablations, and held-out permutation importance.

Results: Each window included 207 stress-event observations and 207 controls. The 5-minute random forest achieved pooled ROC AUC 0.759, average precision 0.772, accuracy 0.686, and F1 0.688. Participant-mean AUC was 0.774 (95% CI 0.694-0.851); all 11 participant AUCs exceeded 0.50. Movement plus temperature achieved participant-mean AUC 0.777, similar to all core signals (0.774).

Conclusions: Short-horizon wearable changes distinguished periods immediately preceding self-reported stress events in this small AUD-recovery sample. These findings do not establish prospective stress-onset, craving, alcohol-use, or relapse prediction.

Keywords: alcohol use disorder; wearable sensors; stress; machine learning; participant-independent validation; digital biomarkers

Introduction

Stress is clinically relevant to substance use disorder recovery. Stress-related changes in subjective, physiological, and neurobiological systems are associated with craving, continued substance use, and relapse vulnerability, although stress alone does not determine whether an individual will return to use [1]. Prospective AUD treatment research has also found that craving can mediate the association between daily stress and next-day alcohol intake [2]. These findings motivate objective monitoring of stress-related states while underscoring that stress detection is not equivalent to substance-use prediction.

Wrist-worn sensors can continuously measure heart rate (HR), electrodermal activity (EDA), skin temperature, and movement in daily life. Prior research has shown that wearable features can differentiate stress- and craving-related episodes among individuals receiving substance use disorder treatment [3]. However, free-living physiological signals are influenced by physical activity, environmental conditions, device contact, circadian patterns, medication, and individual differences. Performance can also be inflated when observations from the same participant appear in both training and testing data. Participant-level separation is therefore essential when evaluating generalization to an unseen wearer [4,5].

The ADARP study collected Empatica E4 wearable signals and ecological momentary assessments from adults with AUD in an ambulatory treatment setting. The original investigations established the feasibility of physiological monitoring and reported associations between several sensor features and self-reported outcomes [6,7]. The dataset also contains participant-generated markers for moments when participants felt more stressed, overwhelmed, or anxious than usual. Whereas the source paper examined a broad period surrounding each marker [7], the present analysis used only pre-marker data and compared several shorter windows using an identical event set.

The primary research question was whether changes from an individual's recent physiological baseline could distinguish periods before self-reported stress events from matched control periods in adults recovering from AUD. Secondary objectives were to determine which sensor modalities and pre-event durations generalized best to unseen participants. Shorter windows were expected to preserve changes near the marker, and multimodal models were expected to outperform single-modality models.

Methods

Study Design and Reporting Framework

This study was a secondary observational machine-learning analysis of deidentified, publicly available ADARP data (Zenodo version 1.0; DOI 10.5281/zenodo.6640290) [8]. Reporting was informed by STROBE and TRIPOD+AI [9,10]. Because the analysis compared matched event and control periods rather than estimating absolute clinical risk, TRIPOD+AI items concerning calibrated individual risk were not fully applicable.

Source Study, Setting, and Participants

The original study recruited adults receiving treatment for mental health conditions and AUD at a treatment agency in Washington State during 2019-2020 [7,8]. Participants had to be 18-65 years old, report at least four standard drinks on five or more occasions in the previous 30 days, have moderate-to-severe AUD under DSM-5 criteria, and own a smartphone. Twenty people were screened, 12 enrolled, and 11 completed the study. Ten of the 11 were women. Participants took part for an average of 14 days, with a range of 7-17 days. The original researchers obtained written informed consent [7].

Participants were instructed to wear the E4 during waking hours, remove it during sleep and water exposure, and keep it on the same wrist. The dataset contains approximately 1,698 hours of recordings, averaging 11.5 hours per participant-day [7]. All 11 completers were included. No a priori sample-size calculation was performed because all eligible observations in the public dataset were used.

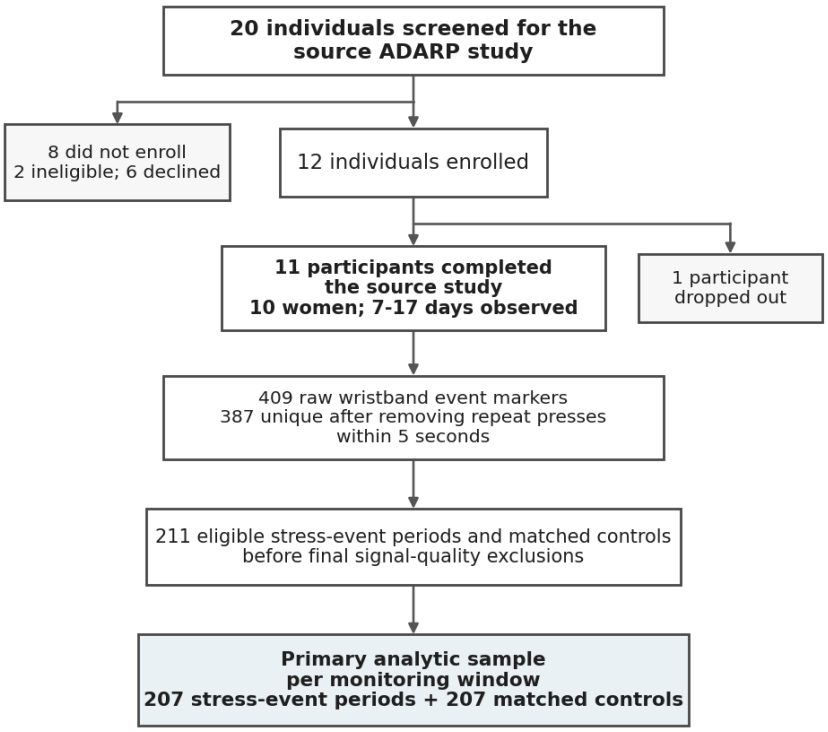


Figure 1. ***Participant and analytic-sample flow. Source-study enrollment counts are from Sah et al. [7]. Analytic event counts were produced by the present preprocessing pipeline.***

Wearable Signals and Event Outcome

The E4 recorded three-axis acceleration at 32 Hz, EDA and skin temperature at 4 Hz, blood-volume pulse at 64 Hz, derived HR at 1 Hz, and irregular interbeat intervals [7]. The primary analysis used HR, EDA, skin temperature, and acceleration magnitude. Oxygen saturation (SpO2) was not available.

Participants pressed the E4 event-marker button when they felt more stressed, overwhelmed, or anxious than usual [7]. A marker therefore represents the participant's annotation time rather than an independently observed physiological onset. The archive contained 409 markers; presses within five seconds of a previous marker were treated as duplicates, leaving 387 unique candidates. The primary outcome contrasted periods immediately before retained markers with matched control periods. It was not a craving, alcohol-use, or relapse outcome.

Matched Controls and Eligibility

To preserve an identical event set across window lengths, each event required continuous coverage for the 60-minute pre-event interval and the immediately preceding 60-minute baseline. Candidate control endpoints were selected within participant, required comparable coverage, could not overlap, and had no stress marker during the analyzed interval or within 60 minutes on either side. Controls were frequency-matched 1:1 to eligible events within participant, producing 211 events and 211 controls before final quality exclusions. An unmarked control denotes the absence of a recorded marker, not proof of absent subjective stress.

Baseline Definition and Feature Engineering

For each endpoint and window length, two adjacent intervals were defined: a comparison window immediately before the endpoint and an equal-length baseline immediately preceding it. Thus, the 5-minute analysis compared the five minutes before an endpoint with the preceding five minutes. The procedure was repeated for 15-, 30-, and 60-minute windows. No post-marker data were used.

Three-axis acceleration was converted to vector magnitude. For HR, EDA, temperature, and acceleration magnitude, the mean, standard deviation, median, minimum, maximum, and linear slope were calculated in both intervals. Each comparison-window statistic was expressed relative to its baseline-window value, yielding 24 core predictors. Positive values indicated increases from baseline and negative values indicated decreases.

Missing Data and Quality Control

Regularly sampled core signals were required to contain at least 95% of expected observations. Four 5-minute stress observations failed this criterion. Those events and four participant-matched controls were removed from every window to preserve a common event set. The final dataset contained 207 stress periods and 207 controls per window (414 observations per window; 1,656 rows total). No core predictor contained missing or infinite values after quality control.

Four additional heart-rate-variability (HRV) features had at least one missing value in 16.3% of the original rows, with the greatest missingness at five minutes. HRV was excluded from the primary model and evaluated separately using median imputation estimated only from each training fold.

Models and Validation

Two binary classifiers were compared. Logistic regression used training-fold median imputation, standardization, balanced class weights, and a maximum of 5,000 iterations. The random forest used 300 trees, a maximum depth of 8, a minimum leaf size of 3, balanced class weights, and random seed 2026. Hyperparameters were fixed across folds and windows rather than optimized on held-out participants. Models were implemented using scikit-learn [11].

Generalization to unseen participants was evaluated using leave-one-participant-out (LOPO) validation. In each of 11 folds, all observations from one participant were reserved for testing and the remaining 10 participants formed the training set. Preprocessing parameters were estimated exclusively within each training fold. This design prevents participant leakage, although it remains internal validation within a single small source study.

Performance Measures and Statistical Analysis

The primary performance measure was ROC AUC, which assesses ranking discrimination across thresholds. Average precision, accuracy, and F1 score were secondary. Metrics were calculated from pooled held-out predictions and separately by participant. Participant-mean AUC confidence intervals used 10,000 bootstrap resamples, and paired participant-level bootstrap resampling compared the 5-minute window with longer windows. No threshold was selected for clinical use because matched sampling produced a 50% event prevalence that differs from continuous monitoring.

Permutation importance was calculated on held-out folds by shuffling each predictor 20 times and measuring the decrease in ROC AUC. Sensitivity analyses compared the 24-feature model with movement alone, movement plus temperature, each modality separately, a model without movement, and a 28-feature model including HRV. No subgroup analysis was attempted because the sample was too small.

Ethics, Registration, and Computational Assistance

The source study obtained written informed consent [7]. This secondary analysis used only deidentified public files and involved no participant contact. The analysis was not preregistered and did not have a separate public protocol. ChatGPT assisted with Python code drafting and debugging, interpretation of intermediate outputs, and manuscript editing. The author verified reported values against saved analysis outputs and retains responsibility for the methods, interpretation, and final text.

Results

Analytic Sample

The final primary dataset included 11 participants. At each duration, 207 pre-marker observations were matched to 207 controls within participant. Each observation contained 24 change-from-baseline predictors. Outcome classes were balanced by design, and no duplicate observations were present.

Table 1. ***Participant-independent model performance by pre-event window.***

Window	Model	Accuracy	F1	ROC AUC	Average precision
5 min	Random forest	0.686	0.688	0.759	0.772
5 min	Logistic regression	0.657	0.618	0.681	0.738
15 min	Random forest	0.594	0.584	0.618	0.604
15 min	Logistic regression	0.570	0.529	0.579	0.606
30 min	Random forest	0.563	0.568	0.616	0.616
30 min	Logistic regression	0.582	0.553	0.613	0.613
60 min	Random forest	0.543	0.545	0.549	0.535
60 min	Logistic regression	0.580	0.558	0.642	0.654

Note. Metrics were calculated from predictions generated only when each participant was held out. Classes were balanced by matched sampling; accuracy and F1 should not be interpreted at real-world event prevalence.

Primary Discrimination and Window Comparison

The 5-minute random forest achieved the strongest pooled discrimination: ROC AUC 0.759, average precision 0.772, accuracy 0.686, and F1 score 0.688 (Table 1). The 5-minute logistic regression achieved ROC AUC 0.681. Random forest performance decreased at longer windows, with AUCs of 0.618 at 15 minutes, 0.616 at 30 minutes, and 0.549 at 60 minutes.

Participant-mean random forest AUC was 0.774 at five minutes (95% bootstrap CI 0.694-0.851), 0.628 at 15 minutes (0.571-0.690), 0.634 at 30 minutes (0.546-0.721), and 0.562 at 60 minutes (0.452-0.668) (Figure 2). The mean five-minute advantage was 0.146 over 15 minutes (95% CI 0.103-0.184), 0.140 over 30 minutes (0.086-0.202), and 0.212 over 60 minutes (0.131-0.320). Five minutes was better for 10 of 11 participants versus 15 minutes and all 11 participants versus 30 and 60 minutes.

At five minutes, participant AUCs ranged from 0.520 to 0.945 (mean 0.774; SD 0.138). All exceeded 0.50, but the range indicates substantial between-participant heterogeneity and near-chance discrimination for at least one participant.

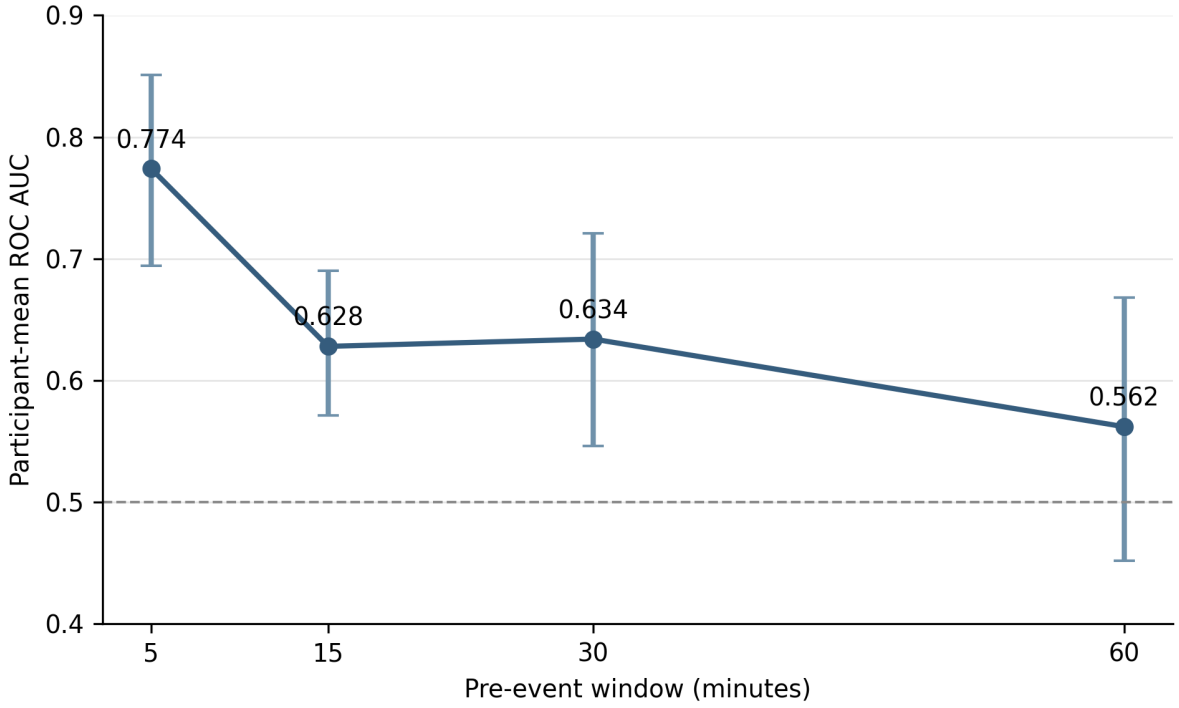


Figure 2. ***Participant-mean random forest ROC AUC by pre-event window. Error bars are 95% confidence intervals from 10,000 participant-level bootstrap resamples; the dashed line is chance discrimination.***

Sensor-Modality Sensitivity Analyses

Movement plus temperature achieved a participant-mean AUC of 0.777 and pooled AUC of 0.760, similar to the full 24-feature model (0.774 and 0.759, respectively). Movement alone achieved 0.744. Excluding movement reduced participant-mean AUC to 0.700, indicating that movement contributed materially but was not the sole source of discrimination. Temperature alone achieved 0.660, whereas EDA, HR, and HRV alone were closer to chance (Figure 3). Adding HRV produced negligible improvement.

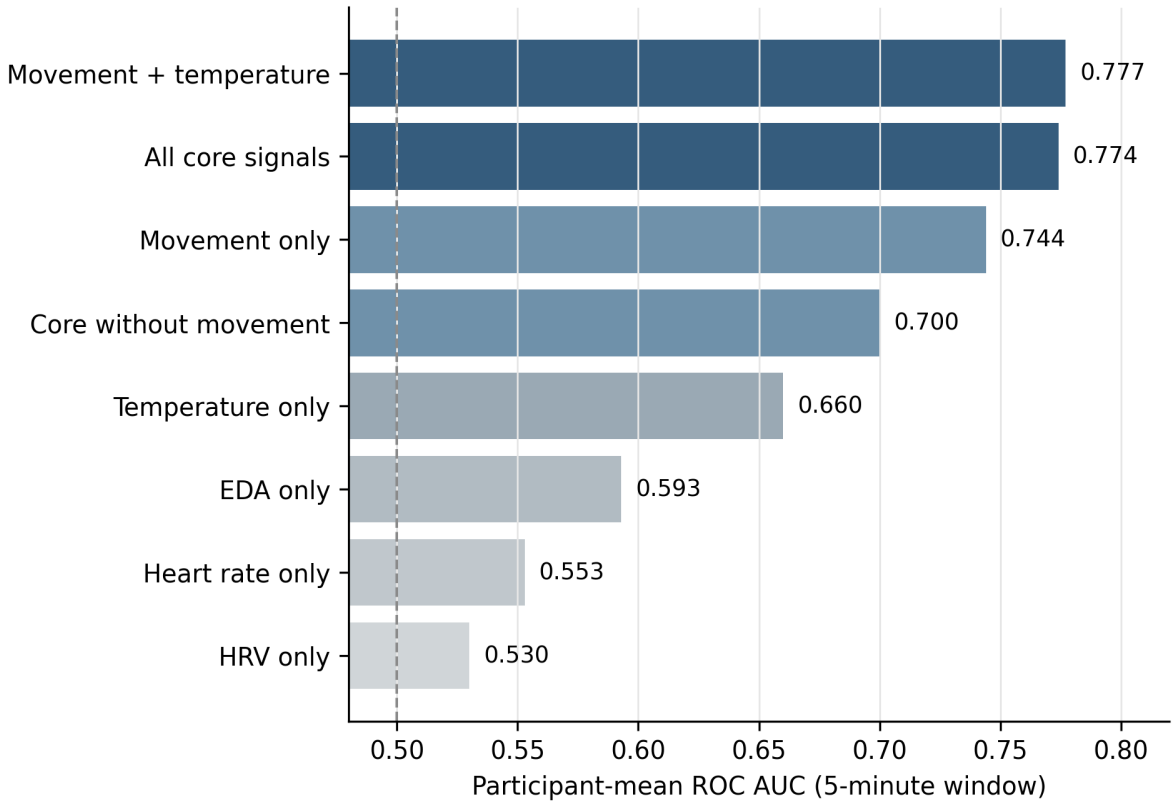


Figure 3. ***Five-minute sensor-modality comparison under leave-one-participant-out validation. Values are participant-mean ROC AUC.***

Feature Importance and Direction

Change in movement variability ranked first in held-out permutation importance (mean AUC decrease 0.0479), followed by temperature variability (0.0448), minimum temperature (0.0242), mean movement (0.0186), and movement slope (0.0175) (Figure 4). These estimates measure model reliance rather than causal physiological effects.

Several patterns appeared across participants. Movement variability was higher before stress markers in 9 of 11 participants, temperature variability was higher in 9 of 11, minimum skin temperature was lower in 10 of 11, and maximum EDA was higher in 9 of 11. These patterns could reflect stress-related arousal, but they could also result from movement, the environment, or changes in how the device contacted the skin.

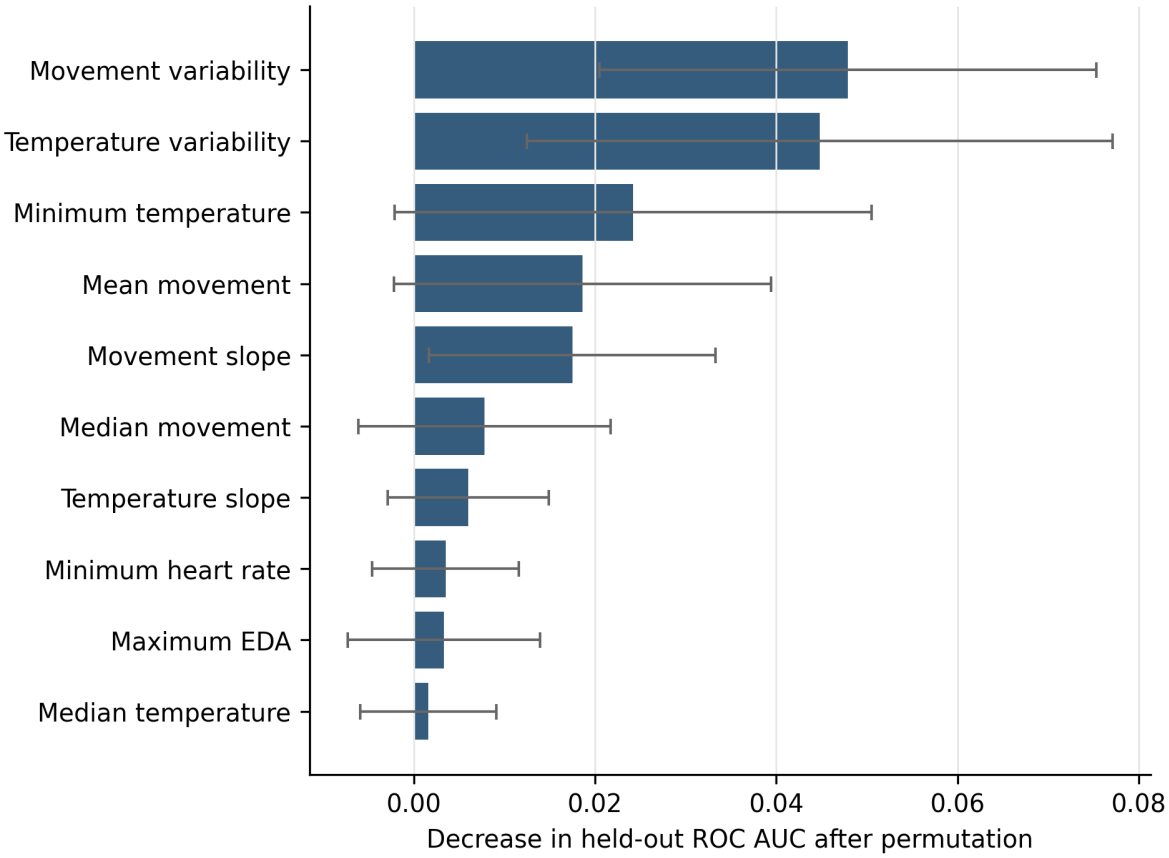


Figure 4. ***Top ten five-minute predictors by mean permutation importance on held-out participants. Error bars show between-participant standard deviation.***

Discussion

Principal Findings

Baseline-relative wearable features distinguished periods immediately before self-reported stress events from matched control periods when evaluated on participants excluded from model training. The 5-minute random forest produced moderate discrimination, with AUC above chance for all 11 held-out participants. Short windows outperformed longer windows, suggesting that detectable differences were localized near the button press rather than sustained across the preceding hour.

Movement and skin temperature provided most of the discriminative information. A 12-feature movement-plus-temperature model matched the full 24-feature model; however, this does not establish a stress-specific biomarker. In free-living settings, physical activity, ambient temperature, device fit, and behavioral context can alter multiple sensor streams simultaneously.

Relation to Prior Research

This analysis extends prior ADARP research by using exclusively pre-marker data, an adjacent within-person baseline, four explicit durations applied to a common event set, four sensor modalities, and participant-level holdout validation [6,7]. The results are consistent with prior substance use disorder studies reporting that physiological data can differentiate stress- and craving-related episodes [3,12]. The lower AUC relative to some previous reports may reflect stricter participant-independent validation, free-living confounding, and the narrower pre-marker definition.

What the Model Does and Does Not Establish

The model performed retrospective event-window discrimination: within a balanced set of pre-marker and matched control periods, it assigned higher scores to many pre-marker periods. The analysis does not establish when stress began, whether the model detected stress before participant awareness, or whether a clinically acceptable alert threshold exists. Participants necessarily recognized stress and pressed the button before an event was labeled.

The model did not predict craving, alcohol use, or relapse. Because it was trained and evaluated on a balanced set of selected event and control windows, its discrimination should not be interpreted as evidence of operational readiness for continuous monitoring.

Strengths

Strengths include real-world ambulatory data from a clinically relevant population, an explicit adjacent baseline, identical underlying events across durations, participant-matched controls, fixed model settings, complete participant separation during validation, participant-level uncertainty estimates, and held-out feature importance.

Limitations and Sources of Bias

The primary limitation is sample size. Eleven participants cannot support stable estimates across demographic or clinical subgroups, and 10 participants were women. Confidence intervals quantify variation within this sample but do not establish transportability to other treatment settings, substances, devices, or populations. The single source study further limits external validity.

Outcome misclassification is also likely. Button presses were subjective point annotations, may have occurred after stress onset, and may have been used inconsistently. Controls may contain unreported stress. Uncontrolled physical activity and environmental temperature could confound movement, HR, EDA, and skin temperature. LOPO validation

prevents participant leakage but does not make repeated observations within participant biologically independent.

The analysis was not preregistered, and several windows, models, and feature sets were examined before the primary specification was finalized. The findings should therefore be considered hypothesis-generating. Because the model was trained on matched pre-marker windows rather than a continuous time series, it cannot estimate real-world alert burden or predict craving, alcohol use, or relapse. Calibration was not evaluated because matched sampling fixed event prevalence at 50%. Validation occurred within one dataset, with no external or prospective evaluation.

Implications and Next Study

These findings support wearable stress-state sensing as a candidate component of a broader recovery-monitoring system, not as a stand-alone relapse alarm. Future research should prospectively collect time-stamped stress, craving, sleep, activity context, intervention delivery, verified substance-use outcomes, and treatment engagement in a larger and more diverse cohort. Prediction targets and alert policies should be prespecified. Evaluation should include participant-held-out and chronological validation, calibration, event sensitivity, false alerts per day, alert precision, and lead time.

Conclusion

Short-term changes in movement, skin temperature, HR, and EDA distinguished periods immediately preceding self-reported stress events from matched control periods in 11 adults receiving AUD treatment. The 5-minute participant-independent random forest achieved moderate discrimination and outperformed longer-window models; movement and temperature contained most of the useful signal. These findings provide proof of concept for event-window stress-state discrimination but do not establish prospective stress-onset, craving, alcohol-use, or relapse prediction. Larger preregistered and externally validated studies are required before clinical use.

Declarations

Data and Code Availability

The source ADARP data are available through Zenodo (https://doi.org/10.5281/zenodo.6640290) [8]. Derived datasets, held-out predictions, and

analysis outputs are retained in the project archive. Analysis code is available from the author upon request and should accompany any public submission to permit exact reproduction.

Funding

This secondary analysis received no dedicated funding. The original ADARP data collection was supported by the Alcohol and Drug Research Program of Washington State University and funds for medical and biological research provided by Washington State Initiative Measure No. 171 [8]. The original funders had no role in this secondary analysis.

Conflicts of Interest

The author declares no conflicts of interest.

Author Contributions

I developed the research question, processed and analyzed the data with computational assistance, interpreted the results, created the figures, and wrote and revised the paper.

Patient and Public Involvement

No patients or members of the public were involved in designing or conducting this secondary analysis. Participants in the source study contributed the data and self-marked event labels.

Use of Artificial Intelligence

I used OpenAI ChatGPT to help draft and debug code, explain intermediate results, plan literature searches, and edit the paper. I checked all reported numbers against the saved analysis files and checked cited claims against the original sources. I take responsibility for the final analysis and writing.

References

1. Sinha R. Stress and substance use disorders: risk, relapse, and treatment outcomes. J Clin Invest. 2024;134(16):e172883. doi:10.1172/JCI172883.

2. Wemm SE, Larkin C, Hermes G, Tennen H, Sinha R. A day-by-day prospective analysis of stress, craving and risk of next day alcohol intake during alcohol use disorder treatment. Drug Alcohol Depend. 2019;204:107569. doi:10.1016/j.drugalcdep.2019.107569.

3. Carreiro S, Chintha KK, Shrestha S, Chapman B, Smelson D, Indic P. Wearable sensor-based detection of stress and craving in patients during treatment for substance use disorder: a mixed methods pilot study. Drug Alcohol Depend. 2020;209:107929. doi:10.1016/j.drugalcdep.2020.107929.

4. Luo W, Phung D, Tran T, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res. 2016;18(12):e323. doi:10.2196/jmir.5870.

5. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns. 2023;4(9):100804. doi:10.1016/j.patter.2023.100804.

6. Alinia P, Sah RK, McDonell M, et al. Associations between physiological signals captured using wearable sensors and self-reported outcomes among adults in alcohol use disorder recovery: development and usability study. JMIR Form Res. 2021;5(7):e27891. doi:10.2196/27891.

7. Sah RK, McDonell M, Pendry P, Parent S, Ghasemzadeh H, Cleveland MJ. ADARP: a multi modal dataset for stress and alcohol relapse quantification in real life setting. In: 2022 IEEE-EMBS International Conference on Wearable and Implantable Body Sensor Networks. 2022. doi:10.1109/BSN56160.2022.9928495.

8. Sah RK, McDonell M, Pendry P, Parent S, Ghasemzadeh H, Cleveland MJ. Alcohol and Drug Abuse Research Program (ADARP) Dataset. Zenodo. Version 1.0. 2022. doi:10.5281/zenodo.6640290.

9. von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement. Ann Intern Med. 2007;147(8):573-577. doi:10.7326/0003-4819-147-8-200710160-00010.

10. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. doi:10.1136/bmj-2023-078378.

11. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825-2830.

12. Shrestha S, Stapp J, Taylor M, Leach R, Carreiro S, Indic P. Towards device agnostic detection of stress and craving in patients with substance use disorder. Proc Annu Hawaii Int Conf Syst Sci. 2023:3156-3163. PMID:36788990.

13. Sah RK, Cleveland MJ, Ghasemzadeh H. Stress monitoring in free-living environments. IEEE J Biomed Health Inform. 2023;27(12):5699-5709. doi:10.1109/JBHI.2023.3315755.

14. Schuurmans AAT, de Looff P, Nijhof KS, et al. Validity of the Empatica E4 wristband to measure heart rate variability parameters: a comparison to electrocardiography. J Med Syst. 2020;44:190. doi:10.1007/s10916-020-01648-w.

Appendix. Participant-Level Five-Minute Results

Table A1. Five-minute random forest performance for each held-out participant.

Participant	Test observations	Accuracy	F1	ROC AUC
101C	20	0.700	0.750	0.880
102C	58	0.638	0.571	0.675
104C	34	0.765	0.750	0.882
105C	32	0.625	0.647	0.641
106C	32	0.812	0.833	0.945
107C	28	0.714	0.733	0.806
108C	32	0.719	0.727	0.828
109C	42	0.690	0.698	0.757
110C	32	0.500	0.467	0.520
111C	54	0.574	0.610	0.643
112C	50	0.840	0.826	0.934

Note. AUC measures ranking within each held-out participant. Test sample sizes are small and should be interpreted with caution.