Research Article

From Signal Purity to Analytical Robustness: Advancing Real-World Mental State Assessment with Consumer-Grade Wearables

Vrinda Marigowda[1], Arun Sasidharan[2], Varsha Vijayan[1], Fayaz Khan Pathan[1], Aravind G[1], Karthik Raghavendran[1], Dhanushya Sree[1], Aaradhana Venkatesan[1], Karthik Venkataraman[1]

1. Neurostellar Private Limited, India; 2. Centre for Consciousness Studies; Dept of Neurophysiology, National Institute of Mental Health and Neuro Sciences, India

Background and Objectives: Reliability in real-world neurophysiology is often hindered by a "pristine signal" fallacy—the assumption that research-grade data is a prerequisite for capturing mental states. We evaluate a paradigm shift: deriving reliability from architectural robustness rather than environmental interference. This study tests a multimodal EEG/PPG framework designed to leverage feature interactions and temporal structure to maintain performance in uncontrolled, naturalistic settings.

Methods: Signal concordance was first assessed by comparing a dry-electrode wearable (Neurostellar) against laboratory-standard EEG/ECG across 31 sessions using repeated measure correlations and inter-class correlations (ICC). We then implemented a multi-layered analytical framework (Random Forest, Gradient Boosting) to classify cognitive/affective tasks and predict subjective states. This framework integrates short-window (15s) spectral/autonomic features with long-window (45s) cluster sequence complexity (Lempel-Ziv), capturing the holistic "Gestalt" of brain-body dynamics. Robustness was evaluated by comparing the stability of multivariate predictions against individual feature volatility under natural and simulated noise.

Results: Feature-level analysis revealed significant inter-subject variability, with moderate-to-strong concordance (>0.5 correlation/ICC) in over 50% of recordings for key features, while others showed substantial discordance. Despite this individual feature volatility, the multi-feature analytical architecture exhibited superior performance, significantly classifying tasks (z=13.08-13.16, p<0.001). For subjective state prediction, while regression yielded modest absolute variance explained (R2=0.16-

0.17), permutation-based evaluation confirmed these models performed significantly above chance ($p < 0.05$, effect size up to 2.37). Feature importance analysis highlighted a balanced reliance on EEG oscillations (e.g., theta-alpha ratio, gamma-alpha ratio), HRV dynamics, and long-window cluster complexity. Crucially, robustness analysis demonstrated that while individual input features suffered deviations exceeding 1000% under noise, the multivariate model predictions remained stable, confirming that reliability is maintained through the ensemble architecture rather than individual signal fidelity.

Conclusions: These findings demonstrate that the reliability of wearables in naturalistic settings is a product of analytical architecture rather than signal purity. By prioritizing multimodal interactions and temporal transitions over isolated, noise-sensitive metrics, consumer-grade devices can effectively track mental states like 'RelaxedFocus.' This shift from signal-centric to architecture-centric validation provides a scalable foundation for real-world neuroscience and long-term cognitive self-regulation.

Correspondence: papers@team.qeios.com — Qeios will forward to the authors

Introduction

In today's highly stimulating digital environment, maintaining sustained attention and cognitive control has become increasingly difficult. The constant influx of information, particularly through smartphones and social media, has led to reduced focus and increased cognitive load[1][2]. While psychometric scales can capture aspects of these mental states, they often fail to reflect the moment-to-moment nuances of lived experience[3]. Objective imaging techniques like EEG offer valuable insights but are mainly limited to laboratory settings. While wearable neurophysiology promises a bridge to naturalistic monitoring[4][5] and tailored training[6][4], its potential is stifled by a critical oversight in current literature: the assumption that high-fidelity, research-grade signals are a prerequisite for capturing mental states. This goal is frequently unattainable in everyday environments.

The Fallacy of the "Pristine Signal"

While laboratory EEG systems provide high signal fidelity, their restrictive nature limits ecological validity. Conversely, wearable sensors facilitate naturalistic monitoring but are inherently susceptible to

doi.org/10.32388/HDMEQ5.2

increased noise and signal degradation[7]. Current literature frequently validates wearables by demonstrating their convergence with research-grade devices in controlled, noise-attenuated, researcher-guided settings[8]. However, this creates a false impression that lab-inspired metrics and algorithms are directly transferable to the real world. When state-of-the-art AI models—trained on such pristine features—encounter user-induced artifacts and environmental noise, their reliability frequently collapses. Furthermore, most available wearables rely on generalized, "one-size-fits-all" models of EEG frequency bands, which ignore the substantial inter-individual variability in neural and autonomic responses[9].

We advocate for a paradigm shift: the reliability of a wearable system in naturalistic settings should be derived from the robustness of its analytical architecture rather than the absence of environmental interference in the signal. Rather than attempting to perfectly "clean" noisy data, we focus on leveraging the rich, interacting information within multimodal features through a multi-layered feature-engineering framework.

Capturing the "Gestalt" via Multimodal Features

To address these challenges, we introduce a unified multimodal framework utilizing frontal EEG and photoplethysmography (PPG) signals captured via a dry-electrode wearable (Neurostellar, India). EEG primarily reflects cortical oscillatory patterns associated with attention and cognitive workload[10][11], while Heart Rate Variability (HRV) derived from PPG reflects autonomic regulation linked to emotional load and arousal[12]. The combination of these modalities offers a pathway to model cognitive–autonomic coupling associated with mental states[13][14][15] that remains stable even when individual signal channels are compromised. The robustness of this architecture is built on the integration of two distinct types of physiological representations into a single machine-learning pipeline:

- **Short-Window Physiological Metrics:** Traditional oscillatory and nonlinear EEG features and autonomic HRV metrics (derived from PPG) are extracted to capture immediate fluctuations in states like focus, cognitive load, and relaxation[13][14], over shorter windows (e.g., 15s).
- **Long-Window "Gestalt" Features:** To supplement these raw metrics, we introduce a representation inspired by EEG microstate literature[16][17]. By transforming synchronized brain-body features into discrete state clusters, we capture the holistic relationship—the Gestalt—between cortical and

autonomic activity. We then quantify the temporal organization of these clusters through sequence complexity (Lempel-Ziv) and transition probabilities over longer windows (e.g., 45s).

By utilizing both types of information as inputs for tree-based ensemble models (Random Forest and Gradient Boosting), the architecture achieves superior resilience. While a brief artifact might corrupt a short-window spectral feature, the long-window complexity feature remains stable, as the overall pattern of state transitions is less susceptible to localized noise. This enables the models to maintain high predictive accuracy even when individual signal channels are compromised.

Research Objective and Core Contributions

This study evaluates this integrated framework through two primary research questions: First, to what extent do EEG and PPG signals from a dry-electrode wearable remain concordant with lab-standard equipment in unconstrained office and residential environments, and where does this concordance break down? Second, can a framework utilizing both individual multimodal features and long-window cluster complexity accurately differentiate subjective states of focus and relaxation across varied naturalistic tasks, such as Chess, a Sternberg task, and music-induced relaxation?

By addressing these questions, this study establishes a paradigm shift, successfully arguing that reliability in the "wild" is derived from analytical robustness rather than elusive signal purity. We achieve this through a methodological innovation that introduces a multi-layered framework, integrating short-window metrics with long-window temporal "Gestalt" features to stabilize state estimation.

Our findings move beyond simple validation to offer several key contributions:

- **Empirical Proof of Concept:** We provide evidence that multivariate architectures are approximately 150 times more stable than raw input features when subjected to natural and simulated artifacts.
- **Validation of Multimodal Redundancy:** We confirm that the cross-referencing of EEG and PPG data significantly outperforms unimodal models ($p < 0.001$).
- **Real-World Application:** We successfully validate the tracking of directional shifts in subjective 'RelaxedFocus' in uncontrolled residential and office environments.

Ultimately, we demonstrate that by building an analytical architecture designed to embrace real-world complexity, we can create scalable tools that serve as genuine companions in a user's journey toward cognitive performance and long-term self-regulation.

Methods

Study Sample

The target sample size was originally determined as sufficient to detect a moderate-to-strong correlation (r = 0.5) between wearable device signals and gold-standard EEG features. A power analysis (alpha = 0.05, 1-beta = 0.80, two-tailed) indicated this to be a minimum requirement of N = 29 independent recording sessions. Sensitivity analysis indicates that with N = 29 (alpha = 0.05, 1-beta = 0.80, one-tailed), the study is also powered to detect a Minimum Detectable Effect (MDE) of g = 0.24 for the machine learning evaluation. This assessment of whether multimodal EEG-HRV features provide predictive power significantly greater than chance utilised a binomial one-sample power analysis within a permutation-based inference framework. This MDE represents a predictive performance of approximately 0.74 (F1-weighted) against a 0.50 null distribution. To capture the longitudinal and inter-subject variability inherent in real-world wearable data, we recruited 10 subjects for 3 recording sessions each (N = 30 total sessions). To maintain the statistical integrity of this nested design, we employed repeated measures correlation (rmcorr) for feature validation, and all machine learning evaluations utilized a Group K-Fold cross-validation approach, ensuring that data from any single participant never appeared simultaneously in both the training and testing sets.

The sample size calculation was done using G*Power software (version 3.1).

This study was performed in accordance with the Declaration of Helsinki. This human study was approved by the Company's Human Ethics Committee (NIEC001/06/2025). All adult participants provided written informed consent to participate in this study.

Participants included males and females aged 18 to 50 years with normal or corrected-to-normal vision and hearing. All subjects had no history of neurological or psychiatric disorders and were recruited via snowball sampling based on predefined inclusion and exclusion criteria. Following written informed consent, subjects were screened using the General Health Questionnaire-28 (GHQ-28) for general well-being and the Edinburgh Handedness Inventory (EHI) for handedness. During recruitment, not all 10 participants complete the 3-day recordings, and we had to recruit additional subjects for single day sessions, achieving 31 recordings from 15 participants. The demographics of the participants are mentioned in Table 1 and the details regarding number of recordings are mentioned in Table 2.

Sl No	Variables	Value
1	Age (Mean ± SD)	33.2 ± 8.74 years
2	Gender	Male = 8; Female = 7
3	Education (Years)	18.84 ± 3.9
4	GHQ-28 score (Mean ± SD)	1.92 ± 2.3
5	Handedness	All right-handed

Table 1. Demographic Details of Participants (N = 15)

Study Design

The study comprised two distinct phases for each participant. For the signal quality comparison, subjects underwent simultaneous recordings using both the Neurostellar's wearable device and a laboratory-standard portable EEG/ECG device during an extended ten-minute resting-state period (five minutes each for eyes-open and eyes-closed conditions). Following a brief interval, participants were monitored solely with the wearable device across four distinct conditions designed to elicit focus, relaxation, and mental effort. These were three-minute conditions with continuous eye opening: performing a modified Sternberg's working memory task, listening to self-selected soothing music, watching self-selected captivating videos, and playing a computer chess match (if proficient). Each condition was preceded and followed by a two-minute resting-state recording with alternating eye closure and opening (see Figure 1 and Table 2). Each participant completed recordings under all conditions within a single day, with this protocol being replicated over two additional days. Consequently, up to three recordings for each specific condition were obtained from each participant. All the recordings were conducted in regular office room or a home environment rather than controlled lab environment. Subjects maintained a position of their comfort, ensuring the environment remained as natural as possible. At the end of each three-minute condition, participants rated their perceived levels of focus and relaxation on a 10-point Visual Analogue Scale (VAS). They were instructed that a scale of 5 represented their baseline rating before the condition; thus, the VAS rating would give us a rating relative to their immediate baseline state.

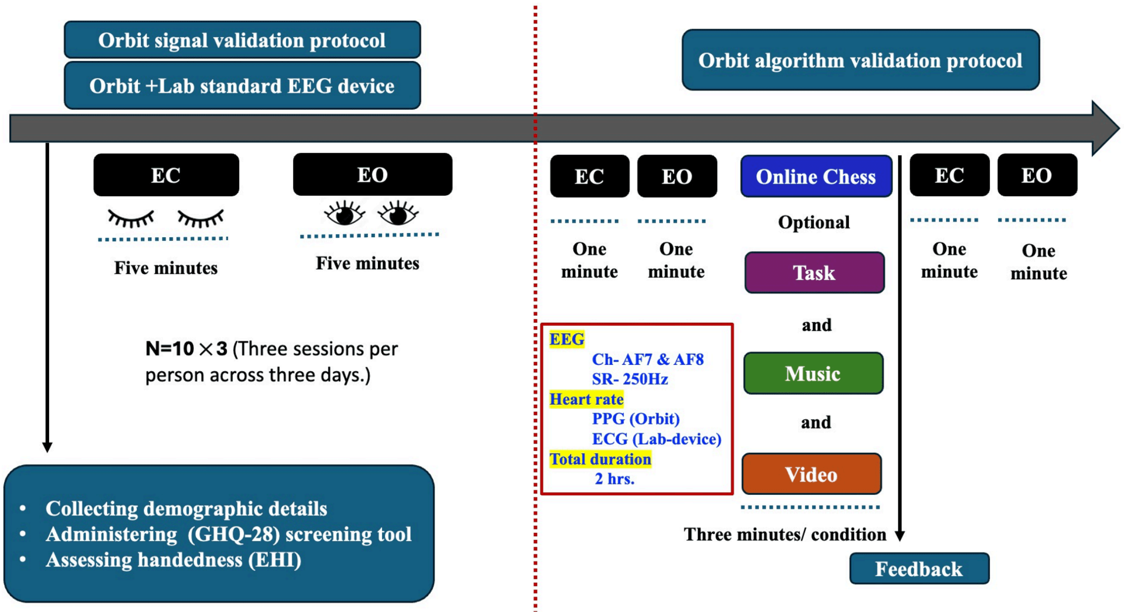


Figure 1. Study Design

The wearable device was self-administered by the subjects, with research staff helping only in instances where hair obstructed the contact between the dry electrodes and the scalp to ensure signal integrity. No skin preparation was performed for the wearable device, which used flat dry electrodes (conductive polymer with Ag-AgCl coating). Conversely, a Lab device (portable battery-operated amplifier; xAMP-L10, Axxonet, India) concurrently recorded two EEG channels at sites proximal to the wearable device's EEG sites (AF7 & AF8), with ground and reference electrodes also in corresponding locations (midline frontal). For the laboratory device, sintered Ag-AgCl ring electrodes were used with electrolyte paste following standard skin preparation, which involved cleansing the sites with 70% isopropyl alcohol to minimize impedance by removing superficial oils and debris. An additional ECG electrode (bottom electrode with ECG gel pad) was placed over the left shoulder. The Lab device was secured in a pouch in front of the abdomen (hung by a neck strap), and electrode cables ran down the neck before connecting to the amplifier. Subjects were in a sitting posture. Both devices recorded EEG at a sampling rate of 250 Hz; ECG was recorded at 250Hz by lab device (Lead 1) and PPG were recorded at 62.5Hz by wearable device (from Fp1 site). Lab device streamed wirelessly (BLE) to a laptop with proprietary acquisition software, which also provided time-locked audio instructions and task markers. The wearable device data was wirelessly streamed to another laptop using custom Python code and saved into a CSV file. To later sync the

doi.org/10.32388/HDMEQ5.2

recordings between the devices, subjects were asked to make a left-right eye movement repeated three times at the beginning of each recording.

Condition	No of Recordings	Device Used
Rest (5 mins Eyes Closed & 5 mins Eyes Open)	**31** (8 subjects x 3 days + 7 subjects x 1 day)	Neurostellar's wearable device and Lab standard Device simultaneously
Playing **Chess** (preRest + Chess + postRest)	**17** (4 subjects x 3 days + 5 subjects x 1 day)	Wearable only
Performing WM **Task** (preRest + Task + postRest)	**30** (8 subjects x 3 days + 6 subjects x 1 day)	Wearable only
Listening to **Music** (preRest + Music + postRest)	**30** (8 subjects x 3 days + 6 subjects x 1 day)	Wearable only
Watching **Video** (preRest + Video + postRest)	**30** (8 subjects x 3 days + 6 subjects x 1 day)	Wearable only

Table 2. Number of recordings

EEG and PPG Feature extraction

The EEG data (AF7 and AF8 channels) underwent basic preprocessing, including 1-40 Hz band-pass filtering (using a finite impulse response (FIR) filter) and 50 Hz notch filtering, via the "MNE-Python" library[18]. The PPG data preprocessing included 1-10 Hz band-pass filtering (FIR filter), 50 Hz notch filtering, a 3-sample moving median filter, and minmax normalisation. Both datasets were segmented into 16-second overlapping epochs with a 0.25-second sliding window.

The study adopted a resilience-based approach rather than aggressive manual artifact rejection. Given that the 16-second overlapping windows with a 0.25-second slide effectively smooth out short-term artifacts, and the amplifier's proximity to the electrodes minimizes movement noise, a high degree of data integrity was maintained. For the signal quality validation, 100% of the recorded segments (31 sessions) were retained for analysis to reflect the raw performance of the device in naturalistic settings. For the classification and regression tasks, only segments with missing values (NaNs), resulting from

doi.org/10.32388/HDMEQ5.2

extreme signal dropouts were filled in using the mean of the respective feature column for that participant, ensuring no entire recordings were discarded.

From EEG epochs, the first principal component analysis (PCA) was selected (this ensured a single time series representing at least 75% of the two-channel data segment), and EEG features were subsequently extracted using functions from multiple Python libraries like "scipy" and "antropy". The 18 EEG features extracted included: power spectral density (welch approach; 4 second window with 50% overlap) in five frequency bands (Delta:1-4Hz, Theta:4-8Hz, Alpha:8-12, Beta:12-30Hz and Gamma:30-40Hz), the power ratios (Theta/Alpha, Theta/Beta, Beta/Alpha and Gamma/Alpha) and non-linear features (Permutation Entropy, Single Value Decomposition Entropy, Sample Entropy, Detrended Fluctuation Analysis, Petrosian Entropy, Katz Fractal Dimension, Higuchi Fractal Dimension and Lempel-Ziv Complexity).

While EEG is inherently a non-stationary signal, the current study utilized 16-second overlapping epochs with a 0.25-second sliding window. This short-time windowing approach effectively captured joint time-frequency components, as the rapid sliding window allows for the tracking of spectral changes over time. By analyzing these quasi-stationary segments, we justify the use of Power Spectral Density (PSD) measures alongside non-linear features to provide a comprehensive characterization of the underlying mental states.

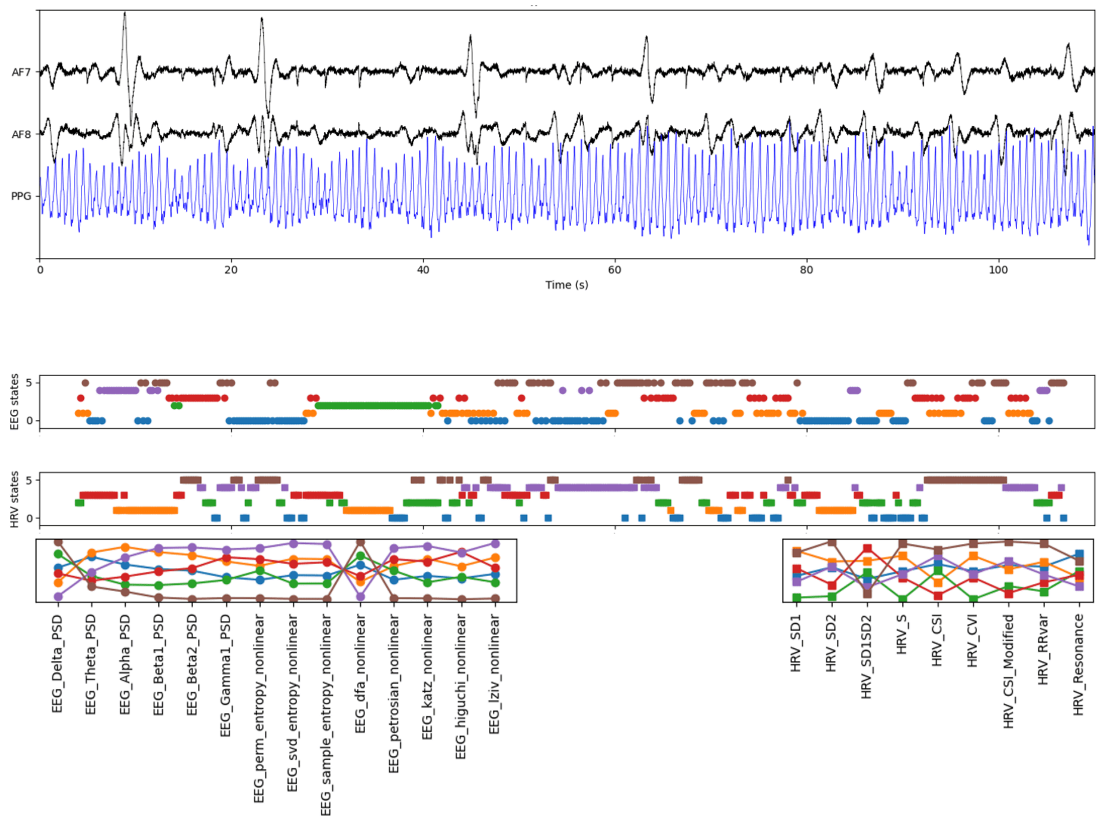
This multifeature selection was initially inspired from the methodology used in sleep staging in a popular sleep stage scoring algorithm using single channel EEG[19] and further expanded with multiple nonlinear measures based on literature citing their relevance in detecting mental states like focus and mind-wandering[20], especially when used in combinations[21]. From the PPG epochs, pulse peak detection[22], correction of erroneous peak placements based on outliers in peak-to-peak differences[23], quality assessment[24], deriving respiratory rate from heart rate[25][26], and extraction of HRV features (here we are using pulse rate variability as a surrogate of conventional HRV) were all done using the "neurokit2" Python library[27]. The 12 HRV features extracted include frequency-domain features (low frequency power [0.04-0.15Hz], high frequency power [0.15-0.4Hz], low frequency / high frequency power ratio and total HRV power), poincaré plot geometry-based non-linear features (like SD1, SD2, SD1/SD2 ratio, ellipse of SD1 and SD2, Cardio Sympathetic Index and Cardio Vagal Index[28][29]) and other statistical measures (like average heart rate, median average deviance of instantaneous heart rate and median average deviance of instantaneous respiratory rate). These short-time window features served as the basis for signal quality correlations between the devices.

Additionally, we extracted a few novel higher-order multi-variate measures (complexity of quasi-stable multi-variate pattern sequences) from the above features derived across multiple consecutive time-windows (every 45s). This analysis utilized a subset of 14 EEG and 9 HRV features to determine six clusters of feature combinations that characterized the entire dataset (Figure 2B), Clustering was performed using an ensemble consensus of 50 k-means models using the scikit-learn library. The 6-cluster framework utilized in this study is theoretically inspired by EEG microstate literature, which characterizes brain activity as a sequence of quasi-stable, functional states. By mapping multimodal features into these discrete clusters, the analysis relies on the 'Gestalt' of the feature set the holistic pattern of brain-body interactions rather than fluctuations in precise raw values. This approach ensures that the resulting cluster sequences were robust to brief environmental noise or localized artifacts, as the overall state assignment remains stable even if individual features experience transient disturbances.

After initial exploration using methods like the Elbow plot, the number of clusters was refined to 6 through an iterative assessment of cluster stability and interpretability, wherein the feature combinations showed meaningful differences, showed consistency across repetitions and subjects, and all clusters represented a good portion of the data (>10%). These 6 cluster patterns were very consistent across different recordings and subjects, and hence a grand mean cluster (clustered from all data) was backfitted to generate cluster sequences (see Figure 2A, bottom panel for grand mean cluster pattern). The resulting sequence of the EEG and HRV cluster dynamics were then subjected to complexity analysis using multiple standard measures (like Higuchi Fractal dimension, Lempel-Ziv Complexity, and Permutation Entropy), as well as a novel measure called Effort-To-Compress (ETC). ETC is defined as the number of iterations needed for a loss-less compression algorithm called Non-Sequential Recursive Pair Substitution (NSRPS) to transform the input sequence to a constant sequence[30]. ETC has been found to be more reliable to capture complexity of short and noisy time series, than traditional entropy measures[31][32]. The complexity of the sequence was calculated across the whole 3-minute recording to get the overall score, and in short sliding windows (45s window with 1s sliding) to track the trend within a setting (Figure 2C). This sequence-based approach can be robust to poor signal quality or artefacts, as it relies on relations between features and not on the intensity of any single feature. For both EEG and HRV cluster sequences, this generated 12 features, which, when added to the individual features, made 45 multi-modal features available for every 45s segment.

doi.org/10.32388/HDMEQ5.2

A


AF7
AF8
PPG
Time (s)
EEG states
HRV states
EEG_Delta_PSD
EEG_Theta_PSD
EEG_Alpha_PSD
EEG_Beta1_PSD
EEG_Beta2_PSD
EEG_Gamma1_PSD
EEG_perm_entropy_nonlinear
EEG_svd_entropy_nonlinear
EEG_sample_entropy_nonlinear
EEG_dfa_nonlinear
EEG_petrosian_nonlinear
EEG_katz_nonlinear
EEG_higuchi_nonlinear
EEG_lziv_nonlinear
HRV_SD1
HRV_SD2
HRV_SD1SD2
HRV_S
HRV_CSI
HRV_CVI
HRV_CSI_Modified
HRV_RRvar
HRV_Resonance

B

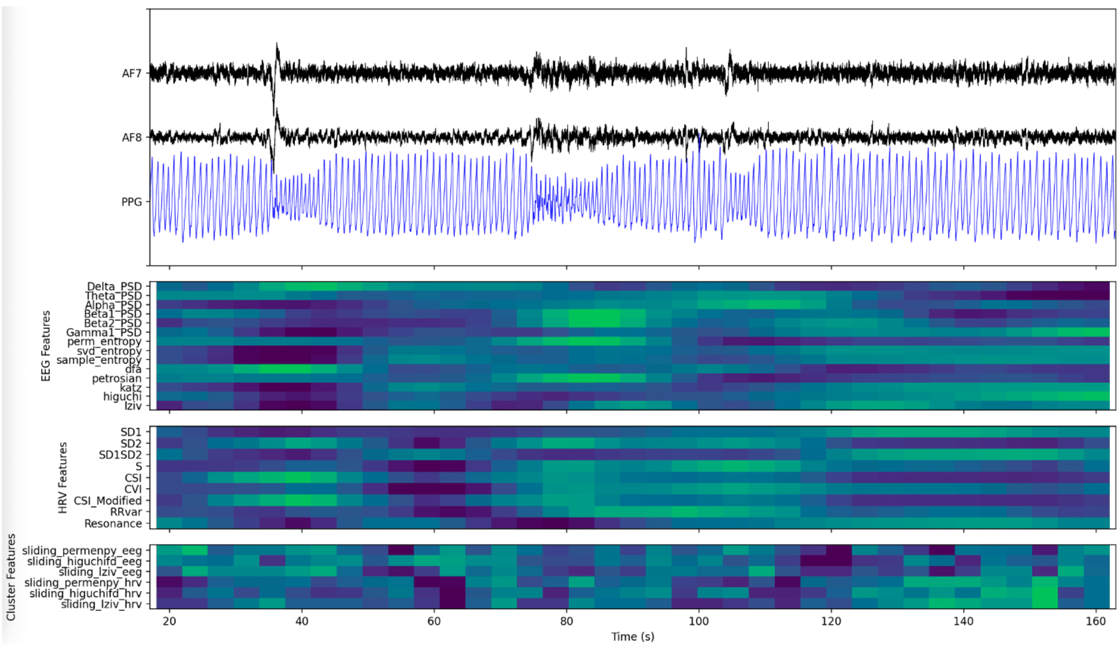


Figure 2. Multimodal Feature extraction from EEG and PPG data. (A) A 110s EEG and PPG segment is converted to quasi-stable multi-variate cluster sequences of the six EEG feature clusters (second panel) and six HRV feature clusters (third panel) based on distinct patterns of 14 EEG features (bottom left panel) and 9 HRV features (bottom right panel). Each colour represents a different cluster (Note that the colouring of EEG and HRV clusters is different). (B) Temporal profiles of the different multimodal features are shown from a data segment. Note that the large amplitude spikes in EEG waveforms or signal drop in PPG waveforms have a differential impact on the different features (including the novel cluster sequences), highlighting the robust nature of this approach. Feature values were z-scored for display (dark colour: low value; light colour: high value).

Machine Learning Analysis

For classification and regression tasks, Random Forest and Gradient Boosting ensemble learners were prioritized over Support Vector Machines or k-Nearest Neighbours. These models provide inherent interpretability through feature importance rankings, facilitating the identification of robust multimodal markers. Furthermore, their recursive partitioning is better suited to capture the high-dimensional, non-linear interactions between EEG and HRV features compared to linear alternatives. Unlike distance-based models, tree-based ensembles are scale-invariant, eliminating the requirement for extensive

doi.org/10.32388/HDMEQ5.2

normalization that could obscure subject-specific physiological variations. The 'sklearn' library was used. The above-mentioned 45 multi-modal features were used in the analysis.

Firstly, as a classification problem, we assessed the ability of the features to predict the condition labels (precondition rest versus that condition, e.g., pre-task rest vs Task). For this, a multivariate permutation testing approach was employed for each participant under each experimental condition. So, for every unique participant and condition combination, the feature set (X) consisted of the 45 features, and the target variable (y) was the two condition labels (precondition rest and the condition). Before model training, infinite values were replaced with Not-a-Number (NaN), and all NaN values were subsequently imputed using the mean of the respective feature column for that participant-condition dataset. A Random Forest Classifier and a Gradient Boost Classifier model with default hyperparameters was used as the predictive model. The predictive performance of the model was evaluated using a 5-fold stratified cross-validation scheme (with data shuffling). The F1-weighted score was used as the primary performance metric, chosen to account for potential class imbalances in the subjective ratings. To determine the statistical significance of the observed F1-weighted score, a permutation test (n_permutations=1000) was conducted. In this procedure, the target variable (y) was randomly shuffled 1000 times, and the predictive model was trained and evaluated (using the same 5-fold stratified cross-validation) on each shuffled dataset. This generated a null distribution of F1-weighted scores. The p-value was calculated as the proportion of permutation scores that were greater than or equal to the actual (unshuffled) F1-weighted score. An effect size for the classification performance was calculated as the difference between the actual F1-weighted score and the mean of the F1-weighted scores obtained from the permuted datasets. This raw effect size was then standardised by dividing it by the standard deviation of the permutation scores, yielding a z-score. This entire procedure—data selection, preprocessing, model training, cross-validation, permutation testing, and effect size calculation—was repeated independently for each participant and each experimental condition, using both models. We also implemented the same approach for single features, to understand the advantage of multi-variate features in terms of effect size.

Next, as a regression problem, we assessed the ability of the features to predict the various subjective ratings ('Focus', 'Relax', and their composite rating 'Relaxed Focus'). As with the classification problem, a multivariate permutation testing approach was independently employed for each subjective rating but combining data from multiple subjects. For each subjective rating (the target variable, y), the corresponding 45 features were used as predictor variables (X). The dataset for each rating was partitioned into a training set (80%) and a testing set (20%) using a stratified split, ensuring all data from one subject stays in one set (train or test). No explicit feature scaling was applied to the training or testing data before the main model training and evaluation, as Random Forest models are generally insensitive to monotonic transformations of features. A Random Forest Regressor or Gradient Boosting Regressor model was employed for predicting each subjective rating. Hyperparameter tuning was conducted on the training data using a grid search approach with 5-fold cross-validation, again using a stratified approach ensuring subject integrity. The hyperparameters explored included n_estimators ([50, 100, 200]) and max_depth ([None, 10, 20]). The model yielding the highest R-squared (R^2) score during cross-validation was selected as the best model. The performance of the optimised model for each subjective rating was assessed on the held-out test set. Evaluation metrics included Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2. Feature importances were also extracted from the trained best model to identify the relative contribution of each objective measure to the prediction. The optimised model for each subjective rating was saved using joblib for persistence. To determine the statistical significance of each optimised model, a permutation test (n_permutations=1000) was performed. This test was conducted on the full dataset (i.e., before train-test splitting) for each subjective rating to assess the overall relationship between the objective measures and the specific rating. The permutation test utilised a 5-fold cross-validation strategy (KFold with shuffling, stratified splitting) and R^2 as the scoring metric. The p-value was calculated as the proportion of R^2 scores from permuted datasets that were greater than or equal to the R^2 score obtained from the original (unpermuted) data. An effect size for the permutation test was calculated as the difference between the R^2 score on the original data and the mean R^2 score from the 1000 permuted datasets. This difference was also standardised by dividing it by the standard deviation of the permutation R^2 scores, providing a z-score-like measure of effect size.

Testing Robustness of multivariate approach to transient artefacts

We used a two-pronged approach that analyzes inherent robustness of the above multivariate measures (ML-based rating predictions) against natural multivariate artifacts in the actual data, and cross-validated it with 80 bootstrapped multivariate artifact injection simulations. To find natural multivariate artefacts, the entire actual feature dataset is scanned, fitted a Minimum Covariance Determinant (MCD) to find the "true" physiological covariance, and used Mahalanobis Distance to find naturally occurring >3-Sigma multivariate anomalies. To generate simulated multivariate artefacts, we bootstrapped 80 trials (arbitrarily chosen) across random subject/condition segments, generating 4-8 Sigma scale multivariate noise vectors that perfectly respected the robust covariance of that subject's baseline. We also calculated the "Max Single Feature Deviation" during every artifact (both natural and simulated), to understand how hard the input features themselves were deviating at that exact moment.

Statistical Analysis

Each of the EEG features, HRV features, and the complexity of cluster sequences was averaged for each setting/session and subjected to statistical analysis. Statistical analysis was done in R-software using Robust statistical measures (takes care of outliers in the data and is immune to data distribution assumptions) from the "WLS2" library. One-sample test (one-sample percentile bootstrap test using the one-step M-estimator) was done using the 'onesampb' function, and Robust ANOVA followed by trimmed mean (10% trim) based post-hoc comparisons were performed. Statistical significance was determined using a $p < 0.05$ threshold after Holm-Bonferroni correction, with exact p-values and 95% Confidence Intervals (CIs) reported for all primary analyses. Additionally, R2 values were utilized to assess the predictive strength of the subject-dependent regression models.

Validation and Predictive Modelling Framework: The analytical approach of this study was bifurcated to address distinct research objectives. First, signal quality validation, comparing the wearable's EEG and PPG features against laboratory standards was conducted as an objective, subject-independent analysis. This step evaluated the device's hardware fidelity in naturalistic settings without reliance on participant feedback. Conversely, the regression analysis for mental state assessment (Focus and Relaxation) was designed as a subject-dependent model. This approach accounts for the significant inter-individual variability in physiological markers of mental states by training models to predict personalized 'RelaxedFocus' ratings relative to each participant's unique baseline and subjective experience.

Results

The "Concordance Gap" in Naturalistic Settings

Visual inspection of time-locked EEG and PPG/ECG snapshots revealed a high degree of hardware fidelity between the wearable and the laboratory standard (Figure 3A, 3B & 3G), with some instances where the wearable recordings exhibited increased noise levels (Figure 3C & 3D) and others where laboratory standard recordings were noisier (Figure 3E & 3F). Quantitative feature-level analysis revealed a significant "Concordance Gap", characterized by high hardware-level fidelity alongside extreme feature-level volatility across subjects and conditions.

doi.org/10.32388/HDMEQ5.2

1. Global vs. Per-Subject Discrepancies

The statistical strength of correlations varied drastically across individual features and subjects (Figure 4A-D).

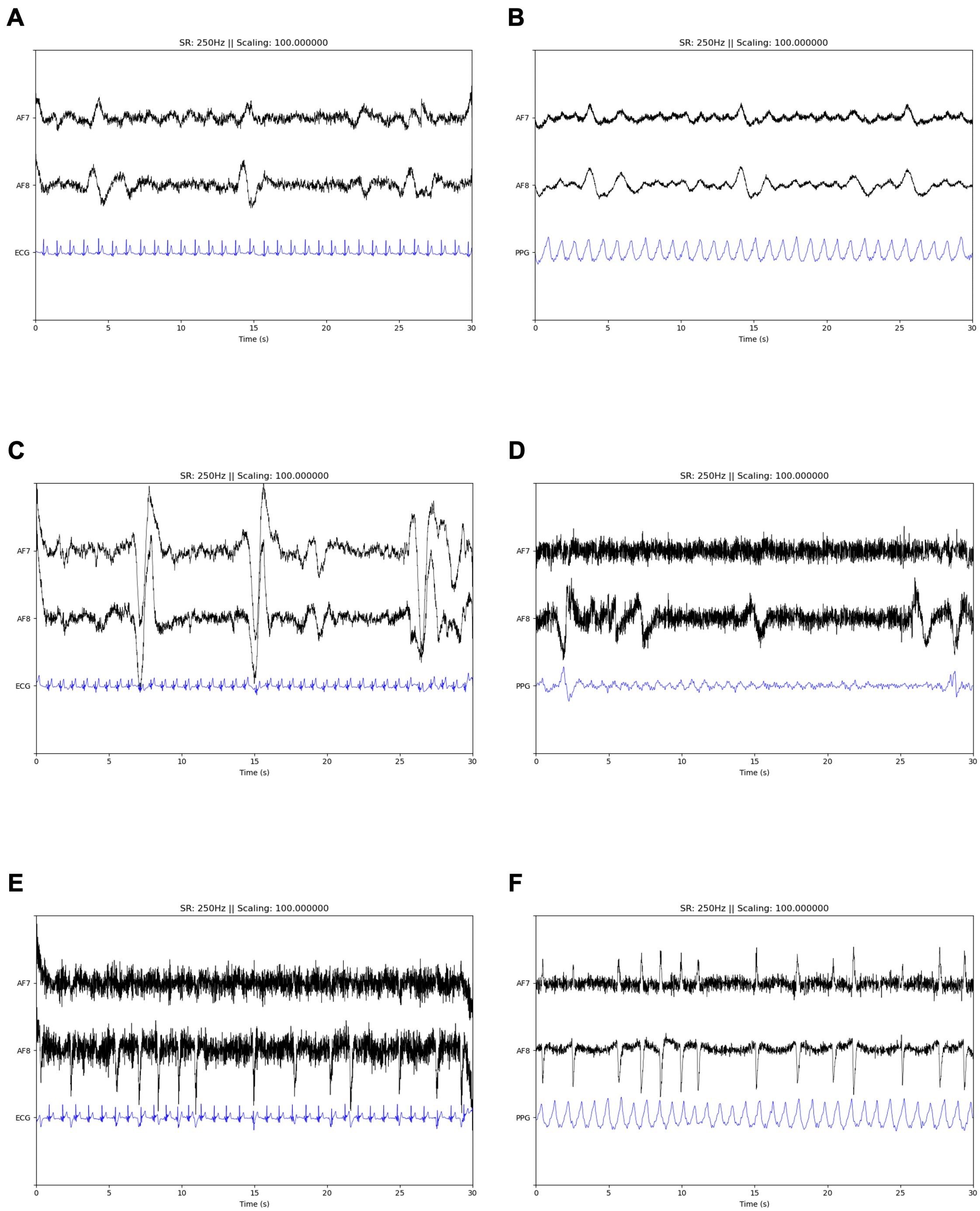
- **Global Underperformance:** Global repeated-measures correlations (RM-Corr) and intra-class correlations (ICC) were consistently lower than the mean of per-subject results, reflecting the destabilizing influence of discordant sessions. For example, while avgHR (Eyes Open) achieved near-perfect correlation in some subjects ($r = 0.999$), its global RM-Corr was only 0.323, illustrating how naturalistic noise in a subset of subjects can mask the device's hardware potential.
- **Feature-Level Volatility:** Individual features exhibited extreme "swings" in reliability. SD2 (Eyes Closed) showed a mean subject correlation of 0.809 (Max: $r = 0.992$), yet other vital HRV markers like hfPower (Eyes Open) fell to a mean r of only 0.194, with some subjects even displaying negative correlations (Min: $r = -0.428$).

2. The Threshold of Reliability

A "Reliability Threshold" analysis (recordings with $r > 0.5$ and $p < 0.05$) highlighted that many standard metrics are inconsistently captured in unconstrained environments (Figure 5A-D):

- **EEG Consistency:** In the Eyes Closed condition, 80% of EEG features (8 of 10 measured, including Alpha, Delta, and Beta bands) maintained moderate-to-high correlation in over half of the recordings. Specifically, TAR (Theta-Alpha Ratio) and GAR (Gamma-Alpha Ratio) were among the most stable, with 87.1% and 77.4% of recordings exceeding the $r > 0.5$ threshold, respectively.
- **HRV Fragility:** Autonomic features were notably more sensitive to recording conditions. While 9 of 12 HRV features met the consistency threshold during Eyes Closed (e.g., SD2 at 83.8%), this dropped to only 6 of 12 features (50%) during Eyes Open sessions. Critical metrics like lfhfRatio achieved significant concordance in fewer than 20% of Eyes Open recordings ($r > 0.5$).

The presence of this "Gap"—where a feature can be highly concordant for one subject but entirely discordant for another—provides empirical proof that unimodal, signal-centric metrics are inherently unreliable for "in-the-wild" applications. These data justify our paradigm shift: by moving away from individual feature magnitude and instead leveraging the multimodal "Gestalt" through tree-based ensemble models, we ensure that the architecture remains robust even when the concordance of individual signal channels collapses.



doi.org/10.32388/HDMEQ5.2

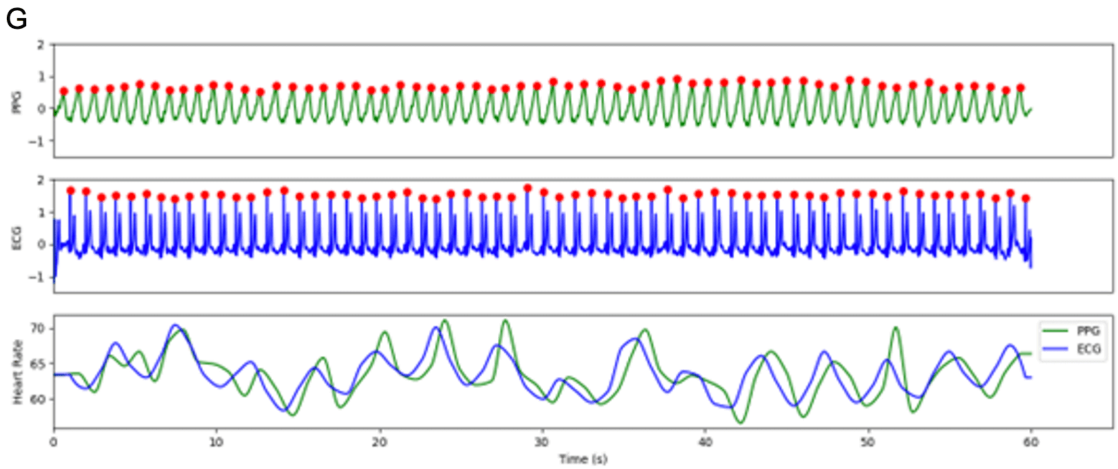
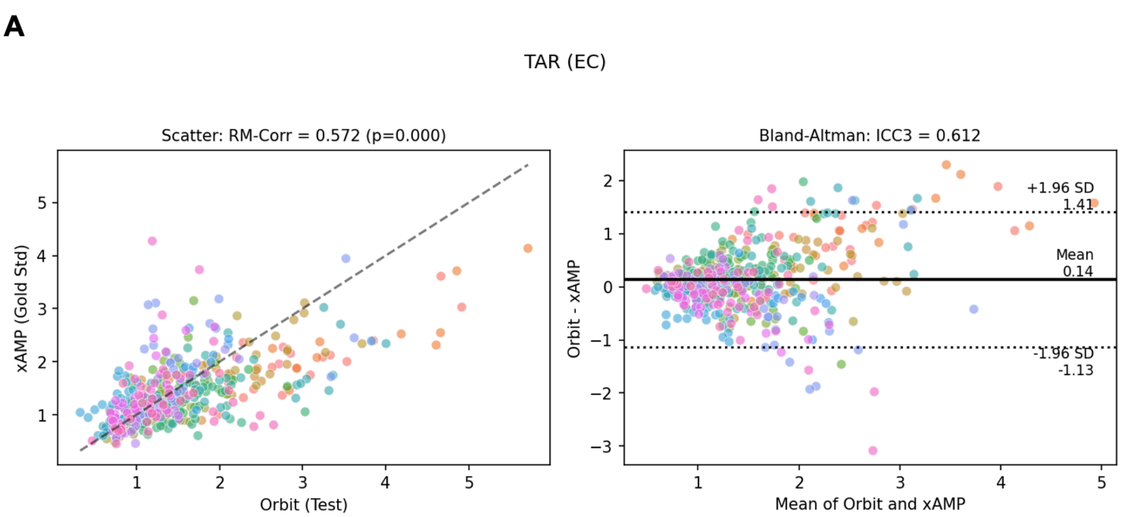


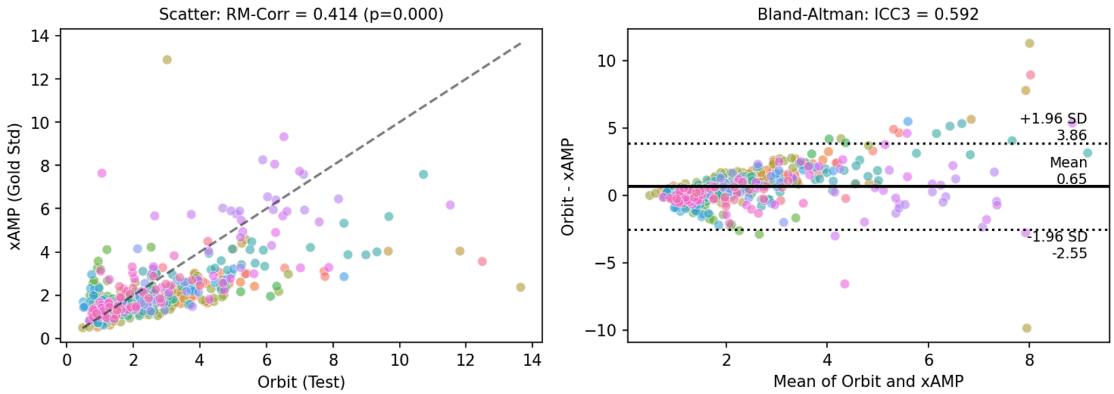
Figure 3. (A-F) 30s snapshots showing time-locked EEG and ECG/PPG pre-processed data from Lab standard device (left) and the wearable device (right), from a representative subject. Some recordings were comparable between the two devices (A & B). Some were noisier in the wearable device (C & D). Some were noisier in the Lab standard device (E & F). (G) Peak detections (red dots) on a 60s data from PPG from the wearable device (top panel; green waveform) and simultaneously acquired ECG from the lab device (middle panel; blue waveform). The instantaneous heart rates in beats-per-minute (bottom panel) are derived from PPG (green) and ECG (blue).



doi.org/10.32388/HDMEQ5.2

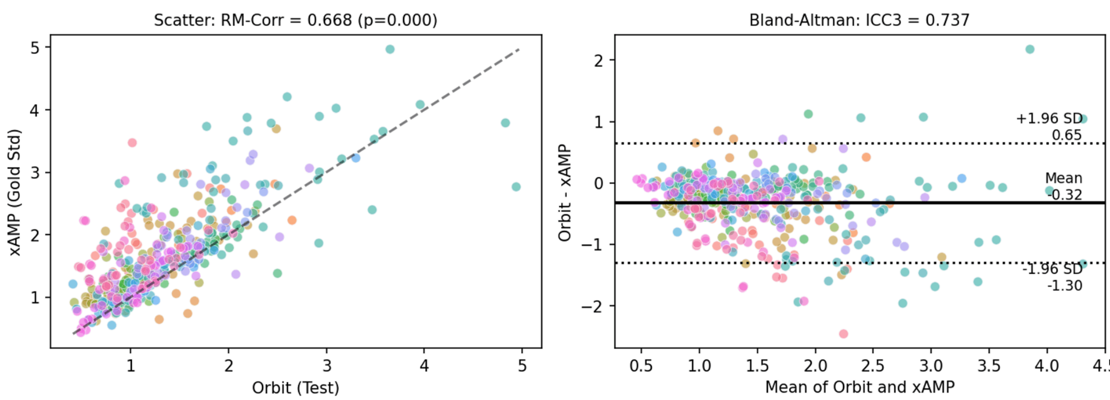
B

TAR (EO)



C

CSI (EC)



doi.org/10.32388/HDMEQ5.2

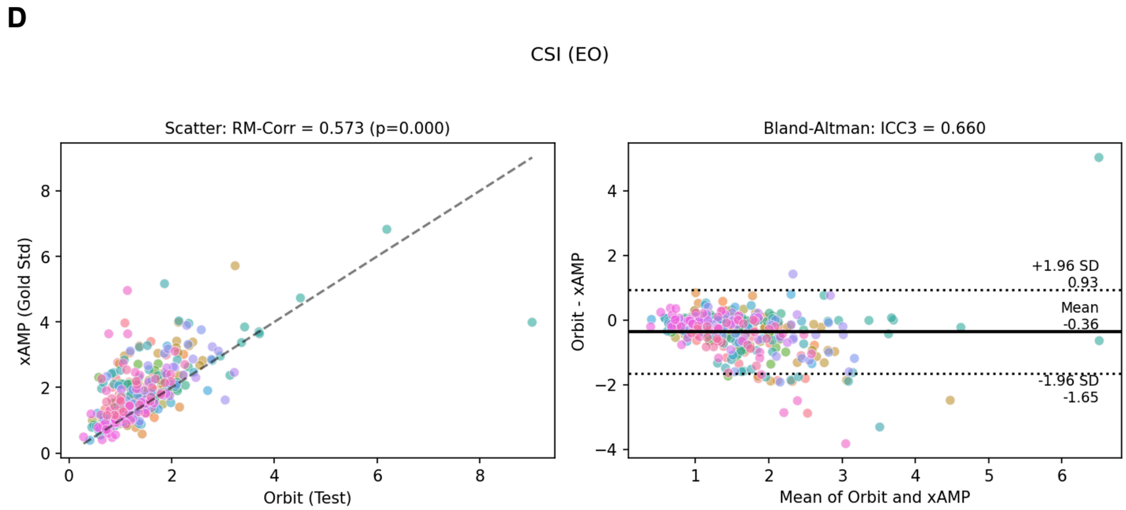
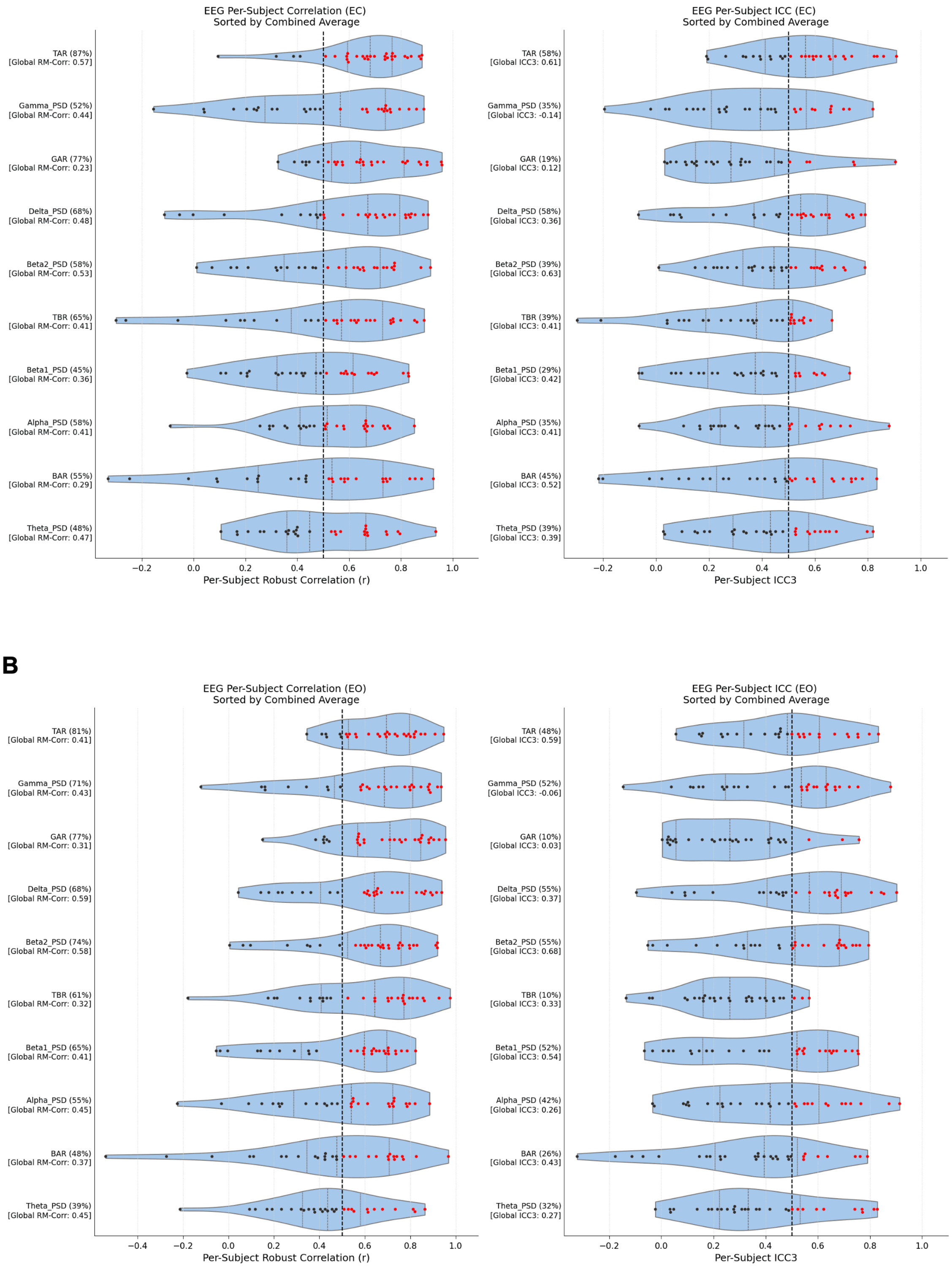


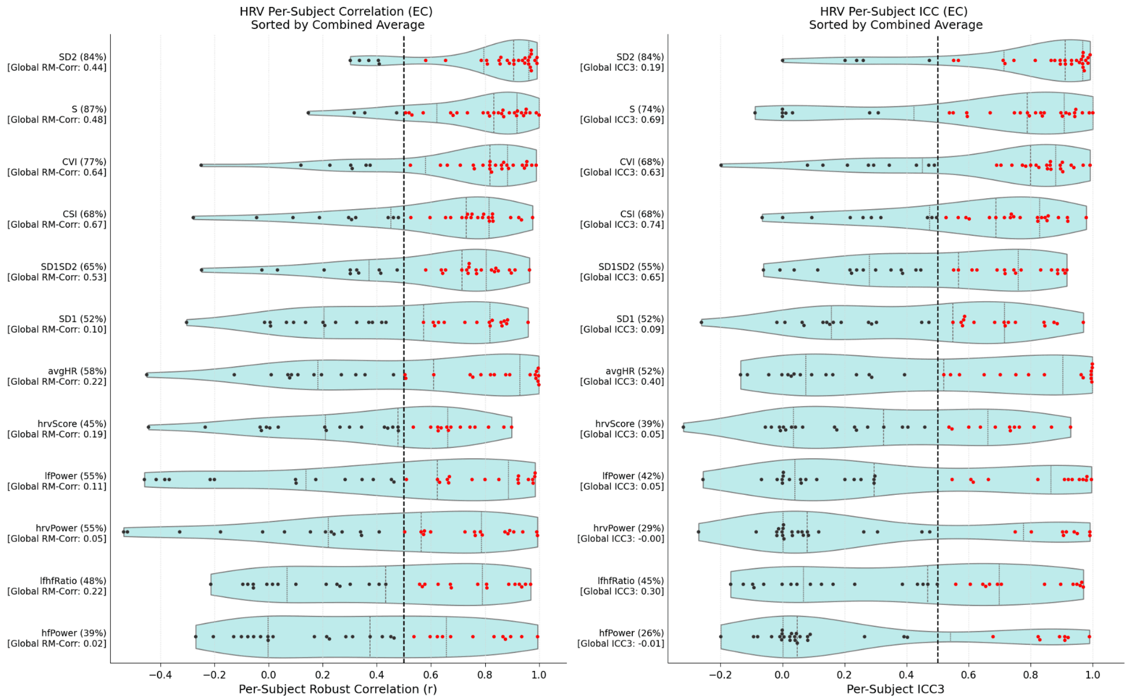
Figure 4. Correlation scatter plots (Left panel) and Bland-Altman plots (Right panel) showing the concordance between xAMP (Lab device) and Orbit (Wearable device) for a selected EEG feature (Theta-Alpha ratio or TAR; A & B) and a selected HRV feature (Cardio-Vagal index or CVI), for each recording. The global repeated-measures correlation (RM-Corr) and intra-class correlation (ICC3) values, across all recordings is shown in the top of the plots. Each dot represents a single 16s epoch (n=31), its colour codes for each subject. This visualizes both the spread of data for an individual subject (within-subject variance) and how different subjects compare to one another (between-subject variance). Data from both eyes closed (EC; A & C) and eyes open conditions (EO; B & D) are shown.

A

B


EEG Per-Subject Correlation (EC)
Sorted by Combined Average
TAR (87%) [Global RM-Corr: 0.57]
Gamma_PSD (52%) [Global RM-Corr: 0.44]
GAR (77%) [Global RM-Corr: 0.23]
Delta_PSD (68%) [Global RM-Corr: 0.48]
Beta2_PSD (58%) [Global RM-Corr: 0.53]
TBR (65%) [Global RM-Corr: 0.41]
Beta1_PSD (45%) [Global RM-Corr: 0.36]
Alpha_PSD (58%) [Global RM-Corr: 0.41]
BAR (55%) [Global RM-Corr: 0.29]
Theta_PSD (48%) [Global RM-Corr: 0.47]
Per-Subject Robust Correlation (r)
EEG Per-Subject ICC (EC)
Sorted by Combined Average
TAR (58%) [Global ICC3: 0.61]
Gamma_PSD (35%) [Global ICC3: -0.14]
GAR (19%) [Global ICC3: 0.12]
Delta_PSD (58%) [Global ICC3: 0.36]
Beta2_PSD (39%) [Global ICC3: 0.63]
TBR (39%) [Global ICC3: 0.41]
Beta1_PSD (29%) [Global ICC3: 0.42]
Alpha_PSD (35%) [Global ICC3: 0.41]
BAR (45%) [Global ICC3: 0.52]
Theta_PSD (39%) [Global ICC3: 0.39]
Per-Subject ICC3
EEG Per-Subject Correlation (EO)
Sorted by Combined Average
TAR (81%) [Global RM-Corr: 0.41]
Gamma_PSD (71%) [Global RM-Corr: 0.43]
GAR (77%) [Global RM-Corr: 0.31]
Delta_PSD (68%) [Global RM-Corr: 0.59]
Beta2_PSD (74%) [Global RM-Corr: 0.58]
TBR (61%) [Global RM-Corr: 0.32]
Beta1_PSD (65%) [Global RM-Corr: 0.41]
Alpha_PSD (55%) [Global RM-Corr: 0.45]
BAR (48%) [Global RM-Corr: 0.37]
Theta_PSD (39%) [Global RM-Corr: 0.45]
Per-Subject Robust Correlation (r)
EEG Per-Subject ICC (EO)
Sorted by Combined Average
TAR (48%) [Global ICC3: 0.59]
Gamma_PSD (52%) [Global ICC3: -0.06]
GAR (10%) [Global ICC3: 0.03]
Delta_PSD (55%) [Global ICC3: 0.37]
Beta2_PSD (55%) [Global ICC3: 0.68]
TBR (10%) [Global ICC3: 0.33]
Beta1_PSD (52%) [Global ICC3: 0.54]
Alpha_PSD (42%) [Global ICC3: 0.26]
BAR (26%) [Global ICC3: 0.43]
Theta_PSD (32%) [Global ICC3: 0.27]
Per-Subject ICC3

c



HRV Per-Subject Correlation (EC)
Sorted by Combined Average
SD2 (84%) [Global RM-Corr: 0.44]
S (87%) [Global RM-Corr: 0.48]
CVI (77%) [Global RM-Corr: 0.64]
CSI (68%) [Global RM-Corr: 0.67]
SD1SD2 (65%) [Global RM-Corr: 0.53]
SD1 (52%) [Global RM-Corr: 0.10]
avgHR (58%) [Global RM-Corr: 0.22]
hrvScore (45%) [Global RM-Corr: 0.19]
lfPower (55%) [Global RM-Corr: 0.11]
hrvPower (55%) [Global RM-Corr: 0.05]
lfhfRatio (48%) [Global RM-Corr: 0.22]
hfPower (39%) [Global RM-Corr: 0.02]
Per-Subject Robust Correlation (r)
HRV Per-Subject ICC (EC)
Sorted by Combined Average
SD2 (84%) [Global ICC3: 0.19]
S (74%) [Global ICC3: 0.69]
CVI (68%) [Global ICC3: 0.63]
CSI (68%) [Global ICC3: 0.74]
SD1SD2 (55%) [Global ICC3: 0.65]
SD1 (52%) [Global ICC3: 0.09]
avgHR (52%) [Global ICC3: 0.40]
hrvScore (39%) [Global ICC3: 0.05]
lfPower (42%) [Global ICC3: 0.05]
hrvPower (29%) [Global ICC3: -0.00]
lfhfRatio (45%) [Global ICC3: 0.30]
hfPower (26%) [Global ICC3: -0.01]
Per-Subject ICC3

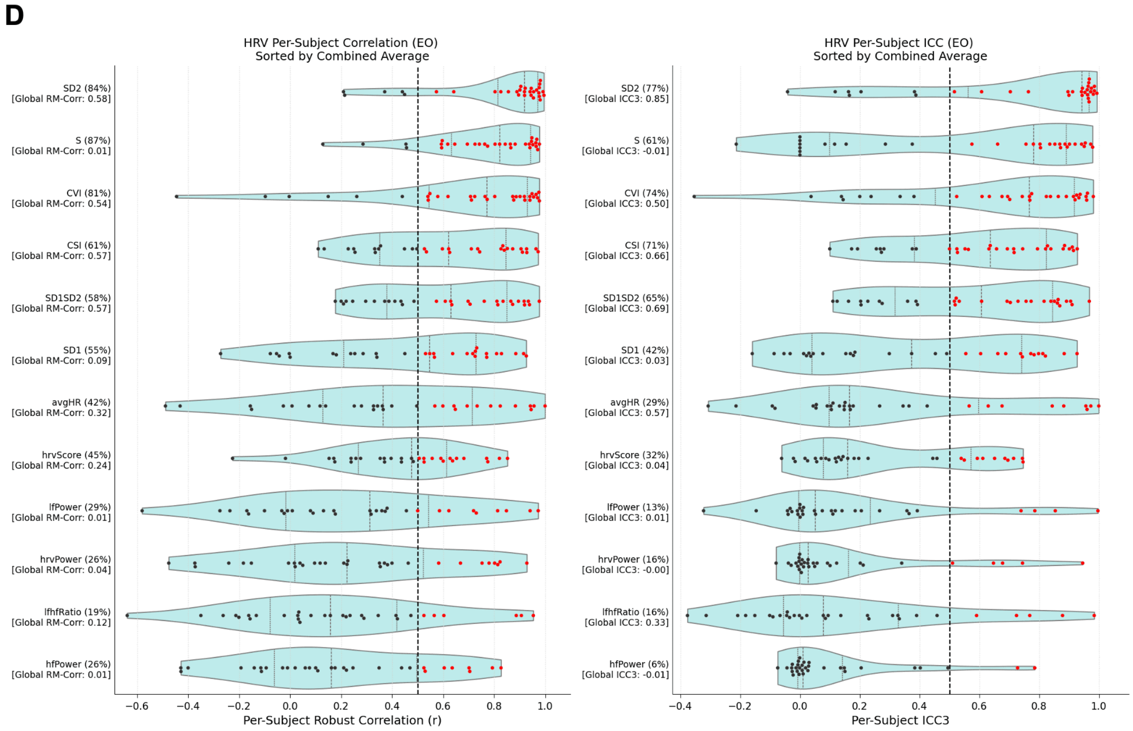


Figure 5. Per-Subject and global-level repeated-measures correlation (RM-Corr) and intra-class correlation (ICC3) values for selected features across all recording data from eyes closed (EC; A & C) and eyes open conditions (EO; B & D). EEG features are shown in the upper panels (A & B), and HRV parameters are in the lower panels (C & D). Each dot represents a single recording (n=31), with red dots representing recordings having moderate-to-high correlations or ICC3 (i.e., r>0.5 & p<0.05) and the percentage values in brackets show the proportion of such recordings, for each feature.

State Classification: Multi-Feature Superiority

The evaluation of machine learning models confirmed that the multi-feature analytical architecture significantly outperformed unimodal approaches in distinguishing mental states across all naturalistic tasks (Chess, Sternberg's task, Music, and Video) (Figure 6).

- **Multi-Feature Stability:** Both Random Forest (RF) and Gradient Boosting (GB) models achieved exceptional effect sizes, with z-scores consistently exceeding 13.00 ($p < 0.001$) for every task. Specifically, the RF model yielded z-scores ranging from 13.08 to 13.11 with minimal variance across subjects.
- **Single-Feature Volatility:** Analysis of individual feature categories revealed significantly lower performance and higher instability.

 doi.org/10.32388/HDMEQ5.2

- **HRV features** achieved a mean z-score of 9.24–9.74 but exhibited a broad range of reliability (Min z = 1.99).
- **EEG Band Power features** showed similar volatility (Mean z = 8.15–9.30), with performance dropping as low as z = 0.21 in some conditions.
- **Cluster Complexity features** (when used as isolated univariate predictors) yielded mean z-scores of 6.59–7.05.

These results demonstrate that the multi-feature framework effectively sifts through discordant individual signals to find stable, task-dependent patterns, achieving a level of consistency unattainable by standard unimodal metrics.

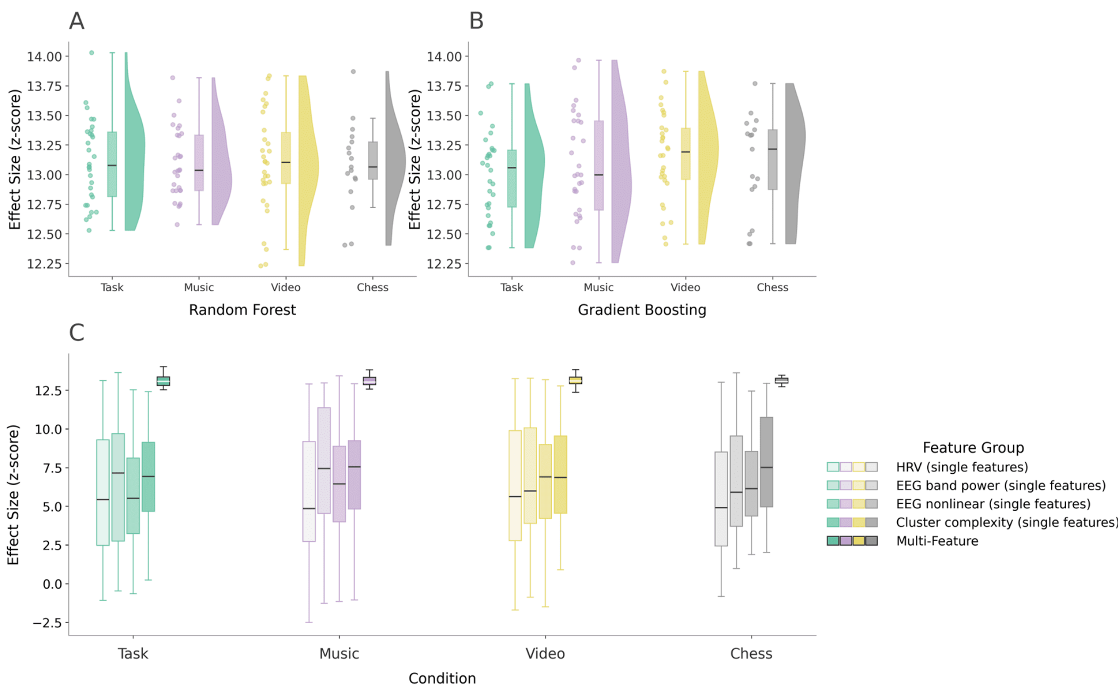


Figure 6. Raincloud Plots showing classification accuracy based on multimodal EEG and HRV features between each condition (Performing Task, Listening to Music, Watching Video or Playing chess) and its pre-condition baseline, using Random Forest Classifier (A) and Gradient Boosting Classifier (B). Each dot represents the multivariate magnitude difference from a single recording computed as a z-score based on a permutation test (1000 permutations); all showed significant change (p<0.001). 106 recordings were used (Task=30, Music=30, Video=30, Chess=17). Box plots showing the large variation in classification accuracy for single features (grouped into categories), when compared to the more consistent pattern for multi-modal feature using Random Forest Classifier is also depicted (C).

 doi.org/10.32388/HDMEQ5.2

Subjective State Prediction and Model Significance

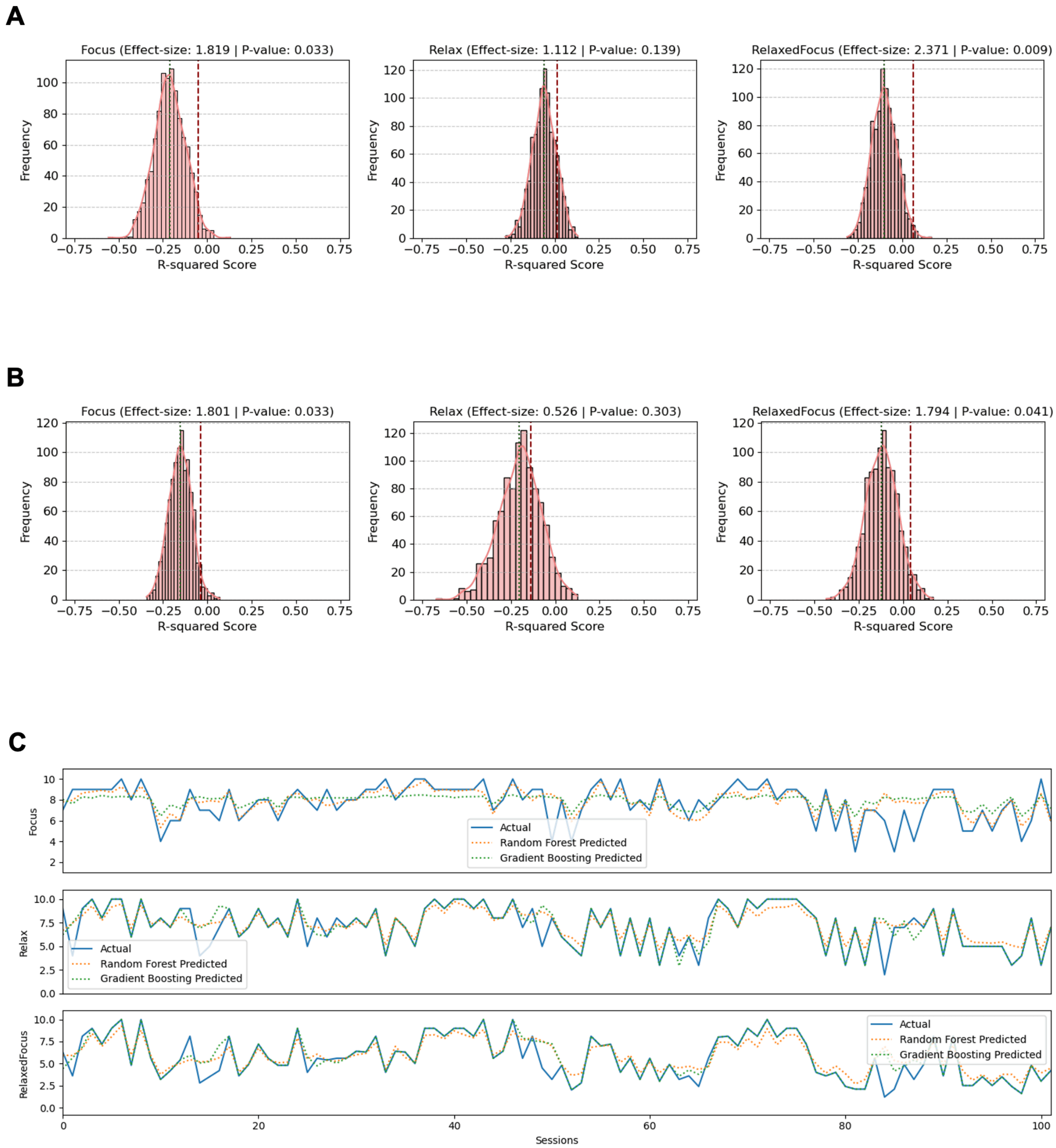
The framework's ability to track nuanced mental states was evaluated via regression of subjective 'Focus', 'Relax', and 'RelaxedFocus' ratings.

- **Ground Truth Validation:** Participants reported statistically significant increases in relaxation during Music ($t = 7.07$, $p < 0.001$) and Video ($t = 8.70$, $p < 0.001$) sessions. Focus ratings increased across all conditions, peaking during goal-directed tasks ($t = 8.48$, $p < 0.001$) compared to passive stimuli like music.
- **Model Significance:** Multivariate permutation tests confirmed that the machine learning models successfully tracked these shifts. While absolute variance explained was modest (R2 = 0.16–0.17) (Figure 7C-D), multivariate permutation tests (1,000 iterations) confirmed that these results were significantly above chance. As shown in the Permutation Histograms (Figure 7A-B), the actual R2 scores (red dotted line) for 'Focus' ($p = 0.033$) and 'RelaxedFocus' ($p = 0.009$) sat well outside the null distribution, providing rigorous statistical evidence of the model's predictive power.
- **Feature Integration:** Predictive success relied on a balanced "Gestalt" of multimodal features. Feature importance rankings (Figure 7F-G) revealed significant contributions from spectral ratios (TAR, GAR), HRV metrics (SD2), and high-level cluster complexity measures (e.g., *sliding_permenpy_eeg*), validating the necessity of an integrated brain-body architecture.

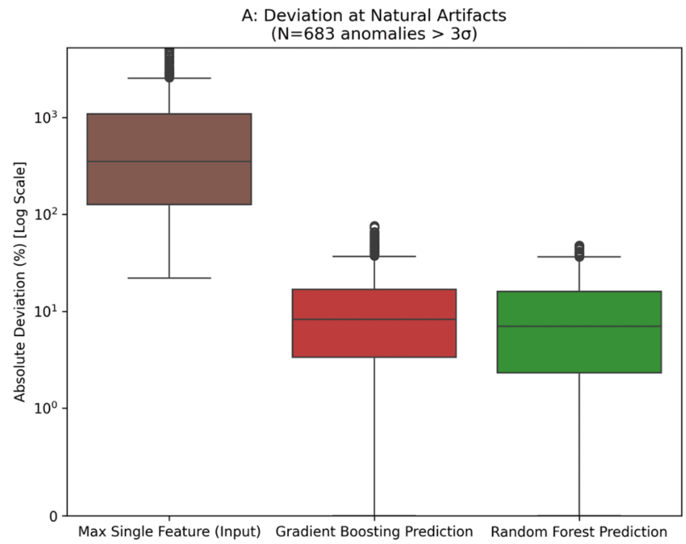
Analysis of the subjective relaxation ratings revealed statistically significant increases during both the Music listening (t-statistic = 7.07, p-value < 0.001) and the Video viewing (t-statistic = 8.70, p-value < 0.001) sessions. These results suggest that both auditory and visual stimuli were effective in inducing a state of relaxation in the participants. Subjective focus ratings also showed statistically significant increases across all experimental conditions: Chess (t-statistic = 8.00, p-value < 0.001), Task (t-statistic = 8.48, p-value < 0.001), Music (t-statistic = 7.30, p-value < 0.001), Video (t-statistic = 8.12, p-value < 0.001). Notably, the Task-oriented session elicited the highest focus ratings, whereas focus was lowest during the Music condition. This suggests that while passive stimuli like music enhance focus to some extent, goal-directed tasks elicited a significantly higher level of mental engagement.

These findings indicate that wearable device data effectively captures meaningful multimodal feature variations for predicting a combined state of Relaxation and Focus.

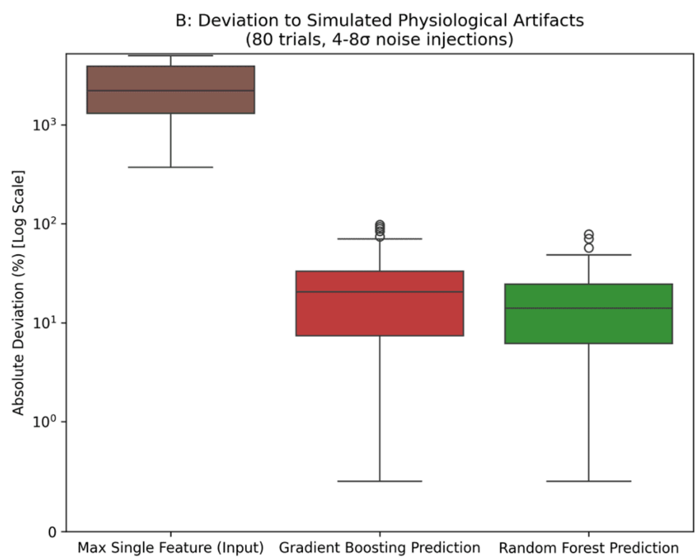
doi.org/10.32388/HDMEQ5.2


A
Focus (Effect-size: 1.819 | P-value: 0.033)
Relax (Effect-size: 1.112 | P-value: 0.139)
RelaxedFocus (Effect-size: 2.371 | P-value: 0.009)
B
Focus (Effect-size: 1.801 | P-value: 0.033)
Relax (Effect-size: 0.526 | P-value: 0.303)
RelaxedFocus (Effect-size: 1.794 | P-value: 0.041)
Frequency
R-squared Score
C
Focus
Relax
RelaxedFocus
Sessions
Actual
Random Forest Predicted
Gradient Boosting Predicted

D

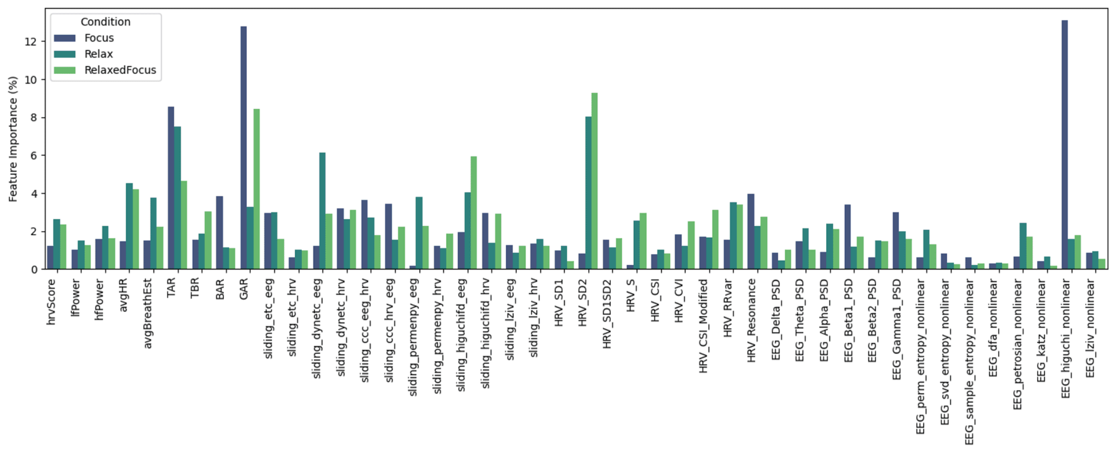


E



doi.org/10.32388/HDMEQ5.2

F



Condition
Focus
Relax
RelaxedFocus
Feature Importance (%)
0
2
4
6
8
10
12
hrvScore
lfPower
hfPower
avgHR
avgBreathEst
TAR
TBR
BAR
GAR
sliding_etc_eeg
sliding_etc_hrv
sliding_dynetc_eeg
sliding_dynetc_hrv
sliding_ccc_eeg_hrv
sliding_ccc_hrv_eeg
sliding_permenpy_eeg
sliding_permenpy_hrv
sliding_higuchifd_eeg
sliding_higuchifd_hrv
sliding_lziv_eeg
sliding_lziv_hrv
HRV_SD1
HRV_SD2
HRV_SD1SD2
HRV_S
HRV_CSI
HRV_CVI
HRV_CSI_Modified
HRV_RRvar
HRV_Resonance
EEG_Delta_PSD
EEG_Theta_PSD
EEG_Alpha_PSD
EEG_Beta1_PSD
EEG_Beta2_PSD
EEG_Gamma1_PSD
EEG_perm_entropy_nonlinear
EEG_svd_entropy_nonlinear
EEG_sample_entropy_nonlinear
EEG_dfa_nonlinear
EEG_petrosian_nonlinear
EEG_katz_nonlinear
EEG_higuchi_nonlinear
EEG_lziv_nonlinear

G

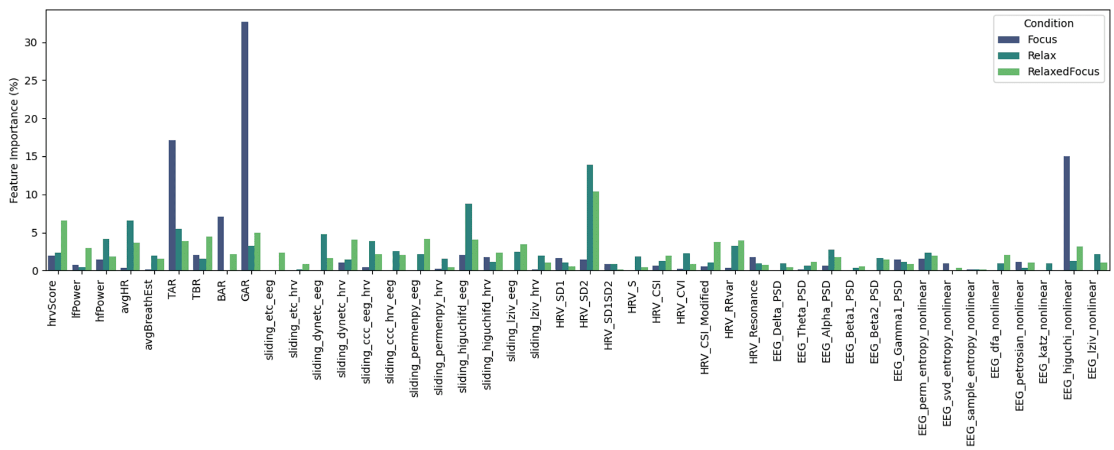


Figure 7. R-squared values showing the performance of the Random Forest (A) and Gradient Boosting (B) Regression models in predicting the subjective Focus, Relax and RelaxedFocus (Focus x Relaxation) ratings, based on multimodal EEG and HRV features from wearable device. The histograms show the distribution of permutation R-squared values (1000 permutations), and the dotted green vertical line represents their mean (chance-level performance of the model). The dotted red vertical line represents the observed R-squared value (actual performance of the model). (C) Line plots showing the variation in Focus, Relax and RelaxedFocus ratings (blue) across the whole 102 recordings used, and how closely the predictions from Random Forest (orange dotted) and Gradient Boosting (green dotted) models match these. (D) Robustness to Natural Artifacts: This plot illustrates that the multivariate measure (RelaxedFocus rating predictions) maintains higher stability than their individual constituent features when subjected to spontaneous, high-sigma anomalies identified via Mahalanobis distance. (E) Robustness to Simulated Artifacts: This plot illustrates that the multivariate measure (RelaxedFocus rating predictions) remains robust even when synthetic perturbations mimicking the interrelated covariance structure of actual biological noise were introduced (an ablation study using 80 randomized, high-magnitude noise injections). (F and G) Bar graphs showing the feature importances of the trained Random Forest and Gradient Boosting regressions models, respectively.

Empirical Demonstration of Architectural Resilience

The primary evidence for a shift toward architecture-centric reliability is found in the model's resilience to both natural and simulated artifacts.

- **Resilience to Natural Artifacts (N=683) (Figure 7D):** In segments where natural multivariate anomalies exceeded 3-sigma, individual input features suffered extreme volatility with a median deviation of 351.00%. In stark contrast, the architecture's final predictions for 'RelaxedFocus' remained remarkably stable, with median deviations of only 6.98% (RF) and 8.28% (GB).

- **Resilience to Simulated Artifacts (N=80) (Figure 7E):** Under extreme conditions with noise injections of 4–8 sigma, the disparity became even more pronounced. While individual features deviated by a median of 2215.05%, the multivariate predictions-maintained stability with median deviations of only 13.96% (RF) and 20.48% (GB).

These data prove that the analytical architecture is 50 to 150 times more stable than its constituent inputs. This validates the core hypothesis: wearable reliability in naturalistic settings is a product of architectural robustness—specifically the integration of multimodal redundancies and temporal complexity—rather than the presence of pristine signals.

Discussion

The current study demonstrates a fundamental paradigm shift for real-world neurophysiology: the reliability of wearable mental state monitoring in naturalistic settings is not a function of signal purity, but a product of analytical architecture. While our hardware validation confirmed a degree of feature-level concordance, the high volatility of these individual signals in uncontrolled environments underscores the limitations of traditional, signal-centric validation[7]. By leveraging a multi-layered framework that integrates multimodal redundancies and temporal complexity, we demonstrate that robust mental-state estimation is achievable despite substantial environmental interference.

The "Concordance Gap": Hardware Capability vs. Feature Fragility

The primary finding of this study is the existence of a "Concordance Gap"—a discrepancy where hardware-level similarity to laboratory standards does not guarantee feature-level stability in naturalistic settings. While the wearable device demonstrated the capacity for high-fidelity recording (e.g., r = 0.99 for *avgHR* in specific sessions), this performance was highly inconsistent across the cohort.

This gap suggests that while the hardware is inherently capable, feature stability is governed by uncontrollable real-world factors rather than equipment quality. Noise is an inherent challenge in electrophysiological recordings, particularly in less controlled, real-world environments, and is not exclusive to consumer wearables[33]. In a controlled lab, factors like skin-electrode impedance, participant movement, and electromagnetic interference (EMI) are minimized. In residential and office settings, however, these variables become dominant. A dry-electrode system may record a "pristine" signal in one session, but minor shifts in headband tension or ambient 50/60Hz noise in another can degrade specific spectral features like *hfPower* or *Theta_PSD* without necessarily ruining the raw signal's visual morphology. Consequently, validation studies restricted to sound-attenuated rooms[8] create a false impression of reliability by ignoring the "wild" variables that lead to this concordance decay. A prior study also reported a similar difference, demonstrating high correlation in frequency domain EEG features (mean=0.957) and acceptable correlation in time domain features (mean=0.580)[34]. A more recent study using in-ear dry electrodes even reported that the correlation with clinical-grade EEG varied based on conditions (0.76 during eyes closed rest and 0.39 during sleep) and overall level of movement and facial artefacts[35].

State Classification: The Empirical Proof of Multi-Feature Superiority

The most striking evidence for our narrative is the superiority of the multi-feature architecture in state classification. While individual features often showed "flashes" of high accuracy, they were inherently volatile across different tasks. For instance, a single EEG band might reliably identify a "Task" state for one subject but fail entirely for another due to the Concordance Gap.

In contrast, our integrated architecture achieved near-perfect classification consistency, with z-scores exceeding 13.00 ($p < 0.001$) across all conditions. This superiority proves that the "whole is greater than the sum of its parts"; the model does not merely average the inputs but extracts a stable "physiological signature" by cross-referencing central (EEG) and peripheral (PPG) data. By outperforming even the best single-feature models, the multi-feature approach demonstrates that the key to real-world utility lies in redundancy—if one signal channel is compromised by the environment, the others provide the necessary context to maintain the state estimate. This is because combining brain activity and autonomic nervous system correlates (HRV) provides a richer, more robust representation of an individual's state than unimodal data alone. Studies have demonstrated the correlation between multimodal brain-computer interface metrics and HRV[36], and have shown that multimodal models incorporating both HRV and electrodermal activity offer more accurate estimates of sympathetic-driven arousal states than unimodal models[37]. The consistency across two different machine learning models further strengthens this conclusion.

The focus should increasingly be on robustly leveraging the rich information within multimodal features and understanding their complex interactions. Prior research already advocates such a paradigm shift in physical activity assessment from wearable accelerometer data[38][39] and mental state assessment from wearable HRV data[40]. Recent evidences also suggest that combining power spectral data and non-linear measures could reflect a deeper understanding of the multifaceted nature of brain function, with oscillations providing invaluable insights into the rhythmic and synchronized activity of local neural populations, while non-linear dynamics offering a powerful window into the global integration and computational complexity of the brain as a whole[41][42]. This approach, acknowledging the inherent noise in real-world data and incorporating the multifaceted nature of brain-body data, enhances resilience to artefacts and allows for the development of models adaptable to individual variability and the nuances of everyday conditions, better capturing subtle mental state shifts.

Architectural Resilience: The 150x Stability Advantage

The most compelling evidence for our narrative is the contrast between input volatility and predictive stability. Our results showed that while natural and simulated artifacts caused individual features to deviate by over 2200%, the multivariate model predictions for 'RelaxedFocus' remained stable with deviations as low as 6.98%.

This stability is achieved through the multimodal "Gestalt." Tree-based ensemble models (Random Forest and Gradient Boosting) are uniquely suited for this task as they do not rely on the absolute magnitude of any single feature, but rather on the hierarchical relationships between them[14]. When a transient artifact corrupts a spectral EEG band, the architecture maintains its "internal logic" by cross-referencing synchronized HRV data and higher-order temporal patterns. This confirms that the system's reliability is derived from the robustness of the analytical architecture rather than the absence of environmental interference.

The Supplementary Role of Cluster Complexity

The inclusion of Long-Window "Gestalt" Features (e.g., Lempel-Ziv complexity of cluster sequences) provided a critical stabilizing layer. Inspired by microstate-based analysis[16][17], this approach prioritizes the temporal organization of states over isolated, noise-sensitive magnitudes.

Our feature importance analysis confirmed the significant contribution of 'Gestalt' markers, which capture the holistic, temporal complexity of brain dynamics (represented as sliding Permutation Entropy [sliding_permenpy_eeg] and Lempel-Ziv Complexity [sliding_lziv_eeg]). This implies that while a 15-second "short-window" feature might be momentarily blinded by a movement artifact, the 45-second "long-window" pattern of state transitions remains a robust marker of the underlying psychophysiological state. This methodology allows the system to embrace, rather than fear, the complexities of real-world data.

Implications: Toward "Cognitive Companions"

This study provides a scalable foundation for a paradigm shift: moving away from the pursuit of elusive signal purity toward robustly leveraging multimodal interactions. For the development of "cognitive companions"[6], this implies that the focus should shift from hardware preparation to architectural resilience.

The successful tracking of subjective 'RelaxedFocus' ($p = 0.009$) suggests that even with modest R2 values, these systems can accurately capture directional shifts in mental states. This makes them ideal for real-world applications such as productivity tracking, stress management, and long-term self-regulation training, where the trend of the state is more important than the absolute precision of a single data point. The subjective experience of "focus" or "relaxation" in isolation may have more idiosyncratic physiological correlates, or their combination may represent a more clearly definable psychophysiological state. Prior studies have also found that the combined use of EEG and HRV features is good for a more definable psychophysiological state, like stress[43], and that higher-order complexity features from EEG differ when predicting states like Focus and Relaxation separately[44].

Limitations and Future Directions

Despite the framework's resilience, several limitations warrant further research.

- **Sample Diversity:** The sample of 15 adults lacks the demographic breadth (age, clinical status) required for universal generalization.
- **Mental Noise:** The modest R2 in subjective regression (0.16–0.17) suggests that "mental noise"—the inherent variance in how individuals perceive and report their internal states—remains a barrier.
- **Fixed Montage:** This study utilized a fixed frontal montage; future work should explore if this architectural resilience holds across different electrode configurations or if parietal/temporal sites introduce unique "Concordance Gaps."
- **Personalization:** Future iterations should explore personalized "Gestalt" models that adapt to a user's unique "noise profile," potentially narrowing the Concordance Gap through individualized feature weighting.

- **Advanced Models:** Future research should prioritise developing and validating advanced machine learning models that learn cross-modal relationships and temporal dependencies, alongside longitudinal studies to assess the evolution and stability of these multimodal signatures.
- **More Devices:** Similar studies should be performed using other lab-grade and wearable devices and on other modalies to strengthen this concept.

Conclusion

This study provides empirical evidence for a necessary shift in real-world neurophysiology: the transition from signal-centric to architecture-centric reliability. Our findings confirm that while wearable hardware can achieve laboratory-standard potential, the inherent "Concordance Gap" in naturalistic environments makes individual physiological features a fragile basis for mental state estimation.

By demonstrating that a multi-layered analytical framework can maintain predictive stability—yielding a 150-fold resilience gain over its volatile constituent inputs—we prove that robust tracking is a product of how data is integrated rather than how it is "cleaned." The superior performance of the multimodal architecture ($z > 13.00$) and the supplementary stability provided by temporal "Gestalt" features offer a scalable blueprint for the next generation of cognitive companions. Ultimately, by embracing the complexity and "noise" of everyday life through robust analytical design, neuro-wearables can move beyond the laboratory to become reliable tools for long-term mental self-regulation and productivity.

Statements and Declarations

Conflicts of Interest

V.M., V.V., P.F.K., A.G., K.R., D.S., A.V., and K.V. are affiliated to Neurostellar. A.S. is neither affiliated to Neurostellar nor received direct any financial compensation from Neurostellar for this study. All authors affirm that the study design, data collection, analysis, interpretation, and manuscript preparation were conducted with scientific objectivity and integrity.

Data Availability

The data required for evaluating the study would be made available by the corresponding author on reasonable request.

Acknowledgements

This study received no external funding. It was supported by Neurostellar, the developer of the wearable device, by providing infrastructure and the device. A.S. arranged the lab-standrard device.

References

1. ^Firth J, Torous J, Stubbs B, Firth JA, Steiner GZ, Smith L, Alvarez-Jimenez M, Gleeson J, Vancampfort D, Armitage CJ, Sarris J (2019). "The "Online Brain": How the Internet May Be Changing Our Cognition." World Psychiatry. **18**(2):119–129. doi:10.1002/wps.20617.
2. ^Shanmugasundaram M, Tamilarasu A (2023). "The Impact of Digital Technology, Social Media, and Artificial Intelligence on Cognitive Functions: A Review." Front Cognit. **2**. doi:10.3389/fcogn.2023.1203077.
3. ^Morgado FFR, Meireles JFF, Neves CM, Amaral ACS, Ferreira MEC (2017). "Scale Development: Ten Main Limitations and Recommendations to Improve Future Research Practices." Psicol Refl Crít. **30**:3. doi:10.1186/s41155-016-0057-1.
4. a, bBorghare PT, Methwani DA, Pathade AG (2024). "A Comprehensive Review on Harnessing Wearable Technology for Enhanced Depression Treatment." Cureus. **16**(8):e66173. doi:10.7759/cureus.66173.
5. ^Tsirmpas C, Andrikopoulos D, Fatouros P, Eleftheriou G, Anguera JA, Kontoangelos K, Papageorgiou C (2022). "Feasibility, Engagement, and Preliminary Clinical Outcomes of a Digital Biodata-Driven Intervention for Anxiety and Depression." Front Digit Health. **4**. doi:10.3389/fdgth.2022.868970.
6. a, bAlhejaili R, Alomainy A (2023). "The Use of Wearable Technology in Providing Assistive Solutions for Mental Well-Being." Sensors. **23**(17):7378. doi:10.3390/s23177378.
7. a, bMakeig S, Kothe C, Mullen T, Bigdely-Shamlo N, Zhang Z, Kreutz-Delgado K (2012). "Evolving Signal Processing for Brain–Computer Interfaces." Proc IEEE. **100**(Special Centennial Issue):1567–1584. doi:10.1109/JPROC.2012.2185009.
8. a, bMikhaylov D, Saeed M, Alhosani MH, Al Wahedi YF (2024). "Comparison of EEG Signal Spectral Characteristics Obtained with Consumer- and Research-Grade Devices." Sensors. **24**(24):8108. doi:10.3390/s24248108.
9. ^Friedman N, Fekete T, Gal K, Shriki O (2019). "EEG-Based Prediction of Cognitive Load in Intelligence Tests." Front Hum Neurosci. **13**. doi:10.3389/fnhum.2019.00191.
10. ^Buzsáki G (2006). Rhythms of the Brain. Oxford University Press. doi:10.1093/acprof:oso/9780195301069.001.0001.

11. ^Cohen MX (2014). Analyzing Neural Time Series Data. MIT Press. doi:10.7551/mitpress/9609.001.0001.

12. ^Shaffer F, Ginsberg JP (2017). "An Overview of Heart Rate Variability Metrics and Norms." Front Public Health. **5**:258. doi:10.3389/fpubh.2017.00258.

13. a, bDi Flumeri G, Giorgi A, Germano D, Ronca V, Vozzi A, Borghini G, Tamborra L, Simonetti I, Capotorto R, Ferrara S, Sciaraffa N, Babiloni F, Aricò P (2023). "A Neuroergonomic Approach Fostered by Wearable EEG for the Multimodal Assessment of Drivers Trainees." Sensors. **23**(20):8389. doi:10.3390/s23208389.

14. a, b, cMark JA, Curtin A, Kraft AE, Ziegler MD, Ayaz H (2024). "Mental Workload Assessment by Monitoring Brain, Heart, and Eye with Six Biomedical Modalities During Six Cognitive Tasks." Front Neuroergonomics. **5**:1345507. doi:10.3389/fnrgo.2024.1345507.

15. ^Mullen TR, Kothe CAE, Chi YM, Ojeda A, Kerth T, Makeig S, Jung T, Cauwenberghs G (2015). "Real-Time Neuroimaging and Cognitive Monitoring Using Wearable Dry EEG." IEEE Trans Biomed Eng. **62**(11):2553–2567. doi:10.1109/TBME.2015.2481482.

16. a, bBréchet L, Michel CM (2022). "EEG Microstates in Altered States of Consciousness." Front Psychol. **13**. doi:10.3389/fpsyg.2022.856697.

17. a, bZhao M, Jia W, Jennings S, Law A, Bourgon A, Su C, Larose M, Grenier H, Bowness D, Zeng Y (2024). "Monitoring Pilot Trainees' Cognitive Control Under a Simulator-Based Training Process with EEG Microstate Analysis." Sci Rep. **14**(1):24632. doi:10.1038/s41598-024-76046-0.

18. ^Gramfort A, Luessi M, Larson E, Engemann DA, Strohmeier D, Brodbeck C, Goj R, Jas M, Brooks T, Parkkonen L, Hämäläinen M (2013). "MEG and EEG Data Analysis with MNE-Python." Front Neurosci. **7**. doi:10.3389/fnins.2013.00267.

19. ^Vallat R, Walker MP (2021). "An Open-Source, High-Performance Tool for Automated Sleep Staging." eLife. **10**:e70092. doi:10.7554/eLife.70092.

20. ^Tang S, Li Z (2024). "EEG Complexity Measures for Detecting Mind Wandering During Video-Based Learning." Sci Rep. **14**(1):8209. doi:10.1038/s41598-024-58889-9.

21. ^Burns T, Rajan R (2015). "Combining Complexity Measures of EEG Data: Multiplying Measures Reveal Previously Hidden Information." F1000Res. **4**:137. doi:10.12688/f1000research.6590.1.

22. ^Elgendi M, Norton I, Brearley M, Abbott D, Schuurmans D (2013). "Systolic Peak Detection in Acceleration Photoplethysmograms Measured from Emergency Responders in Tropical Conditions." PLoS ONE. **8**(10):e76585. doi:10.1371/journal.pone.0076585.

23. ^Lipponen JA, Tarvainen MP (2019). "A Robust Algorithm for Heart Rate Variability Time Series Artefact Correction Using Novel Beat Classification." J Med Eng Technol. **43**(3):173–181. doi:10.1080/03091902.2019.1640

306.

24. ^Orphanidou C, Bonnici T, Charlton P, Clifton D, Vallance D, Tarassenko L (2015). *"Signal Quality Indices for the Electrocardiogram and Photoplethysmogram: Derivation and Applications to Wireless Monitoring." IEEE J Biomed Health Inform.* ***19****(3):1–1. doi:10.1109/JBHI.2014.2338351*.
25. ^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. doi:10.1088/0967-3334/37/4/610*.
26. ^van Gent P, Farah H, van Nes N, van Arem B (2019). *"HeartPy: A Novel Heart Rate Algorithm for the Analysis of Noisy Signals." Transp Res Part F Traffic Psychol Behav.* ***66****:368–378. doi:10.1016/j.trf.2019.09.015*.
27. ^Makowski D, Pham T, Lau ZJ, Brammer JC, Lespinasse F, Pham H, Schölzel C, Chen SHA (2021). *"NeuroKit2: A Python Toolbox for Neurophysiological Signal Processing." Behav Res.* ***53****(4):1689–1696. doi:10.3758/s13428-020-01516-y*.
28. ^Jeppesen J, Beniczky S, Johansen P, Sidenius P, Fuglsang-Frederiksen A (2014). *"Using Lorenz Plot and Cardiac Sympathetic Index of Heart Rate Variability for Detecting Seizures for Patients with Epilepsy." 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 4563–4566. doi:10.1109/EMBC.2014.6944639*.
29. ^Toichi M, Sugiura T, Murai T, Sengoku A (1997). *"A New Method of Assessing Cardiac Autonomic Function and Its Comparison with Spectral Analysis and Coefficient of Variation of R–R Interval." J Auton Nerv Syst.* ***62****(1):79–84. doi:10.1016/S0165-1838(96)00112-9*.
30. ^Nagaraj N, Balasubramanian K, Dey S (2013). *"A New Complexity Measure for Time Series Analysis and Classification." Eur Phys J Spec Top.* ***222****(3):847–860. doi:10.1140/epjst/e2013-01888-9*.
31. ^Kathpalia A, Nagaraj N (2021). *"Time-Reversibility, Causality and Compression-Complexity." Entropy.* ***23****(3):327. doi:10.3390/e23030327*.
32. ^Nagaraj N, Balasubramanian K (2017). *"Dynamical Complexity of Short and Noisy Time Series." Eur Phys J Spec Top.* ***226****(10):2191–2204. doi:10.1140/epjst/e2016-60397-x*.
33. ^Awais MA, Ward T, Redmond P, Healy G (2024). *"From Lab to Life: Assessing the Impact of Real-World Interactions on the Operation of Rapid Serial Visual Presentation-Based Brain-Computer Interfaces." J Neural Eng.* ***21****(4):046011. doi:10.1088/1741-2552/ad5d17*.
34. ^Jackson G, Radhu N, Sun Y, Tallevi K, Ritvo P, Daskalakis ZJ, Grundlehner B, Penders J, Cafazzo JA (2013). *"Comparative Evaluation of an Ambulatory EEG Platform vs. Clinical Gold Standard." 2013 35th Annual Inter*

*national Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). 1222–1225. doi:*10.1109/EMBC.2013.6609727.

35. ^*Pazuelo J, Juez JY, Moumane H, Pyrzowski J, Mayor L, Segura-Quijano FE, Valderrama M, Le Van Quyen M (2024). "Evaluating the Electroencephalographic Signal Quality of an In-Ear Wearable Device." Sensors.* **24***(12):3973. doi:*10.3390/s24123973.
36. ^*Vasilyev V, Borisov V, Syskov A, Kublanov V (2019). "Case Study of Interrelation Between Brain-Computer Interface Based Multimodal Metric and Heart Rate Variability." Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies. 532–538. doi:*10.5220/0007694505320538.
37. ^*Subramanian S, Brown E, Barbieri R (2020). "Multimodal vs Unimodal Estimation of Sympathetic-Driven Arousal States." 2020 Computing in Cardiology Conference. doi:*10.22489/CinC.2020.290.
38. ^*Tiwari A, Cassani R, Kshirsagar S, Tobon DP, Zhu Y, Falk TH (2022). "Modulation Spectral Signal Representation for Quality Measurement and Enhancement of Wearable Device Data: A Technical Note." Sensors.* **22***(12):4579. doi:*10.3390/s22124579.
39. ^*Troiano RP, McClain JJ, Brychta RJ, Chen KY (2014). "Evolution of Accelerometer Methods for Physical Activity Research." Br J Sports Med.* **48***(13):1019–1023. doi:*10.1136/bjsports-2014-093546.
40. ^*Tiwari A, Falk TH (2021). "New Measures of Heart Rate Variability Based on Subband Tachogram Complexity and Spectral Characteristics for Improved Stress and Anxiety Monitoring in Highly Ecological Settings." Front Signal Process.* **1***:737881. doi:*10.3389/frsip.2021.737881.
41. ^*Balkenhol J, Händel B, Biswas S, Grohmann J, Kistowski Jv, Prada J, Bosman CA, Ehrenreich H, Wojcik SM, Kounev S, Blum R, Dandekar T (2024). "Beyond-Local Neural Information Processing in Neuronal Networks." Comput Struct Biotechnol J.* **23***:4288–4305. doi:*10.1016/j.csbj.2024.10.040.
42. ^*Guevara Erra R, Perez Velazquez JL, Rosenblum M (2017). "Neural Synchronization from the Perspective of Non-Linear Dynamics." Front Comput Neurosci.* **11***:98. doi:*10.3389/fncom.2017.00098.
43. ^*Attar ET, Balasubramanian V, Subasi E, Kaya M (2021). "Stress Analysis Based on Simultaneous Heart Rate Variability and EEG Monitoring." IEEE J Transl Eng Health Med.* **9***:1–7. doi:*10.1109/JTEHM.2021.3106803.
44. ^*Richer R, Zhao N, Amores J, Eskofier BM, Paradiso JA (2018). "Real-Time Mental State Recognition Using a Wearable EEG." 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). 5495–5498. doi:*10.1109/EMBC.2018.8513653.

Declarations

Funding: This study received no external funding. It was supported by Neurostellar, the developer of the Orbit device, by providing infrastructure and the Orbit device. A.S. provided the lab-standard device.

Potential competing interests: V.M., V.V., P.F.K., A.G. and K.R. are affiliated to Neurostellar. A.S. is neither affiliated to Neurostellar nor received direct financial compensation from Neurostellar for this study. All authors affirm that the study design, data collection, analysis, interpretation, and manuscript preparation were conducted with scientific objectivity and integrity.