

**Personalized Baseline Models for EEG Stress Detection Under
Wearable-Constrained Conditions**

A pilot study comparing unsupervised subject-specific deviation models against supervised population classifiers on the SAM40 dataset

By Öykü Nur Kesek

Contents

Abstract	3
1. Introduction	3
2. Related Work	4
3. Methods	4
3.1 Dataset	5
3.2 Preprocessing	5
3.3 Feature Extraction: Differential Entropy	6
3.4 Models	6
3.5 Statistical Analysis	7
4. Results	7
4.1 Model Comparison	7
4.2 SHAP Feature Attribution (Population Classifier)	10
4.3 Extended Analyses	12
5. Discussion	13
5.1 Study Limitations	13
6. Conclusion	14
7. Future Directions	15
References	16

Abstract

This study tests whether an unsupervised personal baseline deviation model outperforms a supervised population classifier for EEG stress detection under wearable-constrained conditions, using two temporal electrodes (T7/T8) to simulate consumer behind-ear devices. On the SAM40 dataset (40 subjects, 32-channel EEG, 128 Hz), the personal baseline model significantly outperformed the population classifier in the wearable condition (accuracy: 0.611 vs 0.538, $p=0.025$, $r=0.355$) and full 32-channel condition (accuracy: 0.693 vs 0.619, $p=0.044$, $r=0.318$). SHAP analysis identified temporal alpha differential entropy as the dominant stress biomarker, 2.4x more important than any other band. Alpha suppression occurred in 27/40 subjects; the remaining 13/40 showing enhancement represent a subpopulation for whom directionally rigid models fail. Results constitute an upper bound on real wearable performance as preprocessing used all 32 channels before electrode extraction.

Keywords: wearable EEG, personalization, stress detection, longitudinal variability, differential entropy, brain-computer interface

1. Introduction

Psychological stress is a growing public health concern, and its objective measurement through physiological signals has attracted significant research attention. Electroencephalography (EEG) is an effective tool for identifying stress as it detects the cognitive aspects of stress prior to the emergence of peripheral reactions such as changes in heart rate or changes in skin conductance [1]. Even though EEG wearables have made ambulatory brain monitoring increasingly accessible, two fundamental problems prevent reliable deployment.

First is the inter-subject variability: resting alpha power, peak frequency, and spectral distributions vary significantly among individuals [1, 2]. Population classifiers trained on averaged patterns across many subjects learn a mean that may not represent any individual accurately, causing systematic misclassification in subjects whose baseline deviates from the group mean. Second, the electrode constraint: devices such as the Emotiv MN8 and use only two electrodes. Specifically, Emotiv MN8 uses electrodes that are behind the ear at temporal positions T7 and T8. Detection algorithms developed for 32-channel laboratory systems, however, do not indicate the extent to which performance deteriorates under this constraint.

This study addresses both problems simultaneously. We compare an unsupervised personal baseline deviation model against a supervised gradient-boosting population classifier under T7/T8-restricted and full 32-channel conditions. Previous work has benchmarked cross-subject classifiers on public datasets including DEAP [10], but the specific question of whether personalisation advantage persists under two-channel wearable constraints has not been quantified with proper leave-one-subject-out evaluation. This is Layer 1 of a three-layer research

program: pilot analysis on public data (this study), original data collection with validated stress induction (Layer 2), and real wearable hardware validation (Layer 3).

2. Related Work

Inter-subject EEG variability is documented as the primary obstacle to cross-subject generalisation [1, 2]. Domain adaptation, transfer learning, and normalisation approaches have been proposed as solutions [2, 3]; Fdez et al. [3] demonstrated that standard batch normalisation fails when subjects differ systematically in their baseline distributions, thereby supporting per-subject calibration as a practical alternative. Alpha band suppression (8-13 Hz) is the most consistently reported EEG stress correlate [4, 5]: Packheiser et al. [6] found significant frontal alpha asymmetry changes during TSST-induced stress in 51 healthy adults, and Schapkin et al. [4] identified alpha reduction as the most sensitive indicator of cognitive load. In-ear EEG signals correlate significantly with scalp EEG at T7/T8 positions [7, 8], validating T7/T8 extraction as a methodologically sound simulation of behind-ear wearables, though with the upper-bound caveat detailed in Section 3.2. A 2025 systematic review of 39 low-cost EEG stress studies confirmed alpha suppression as the dominant biomarker and noted that no reviewed study released source code [9], motivating the open-source pipeline presented here. This study asks whether a zero-label, subject-specific baseline can match or surpass the practical performance of a supervised cross-subject classifier under wearable EEG constraints, which is a more deployment-focused question instead of trying to outperform state-of-the-art cross-subject deep learning systems.

3. Methods

3.1 Dataset

The SAM40 dataset [11] contains 32-channel EEG recordings from 40 adult subjects (gender and age distributions are reported in the original publication; dataset available at: <https://www.kaggle.com/datasets/benbenbite/sam40-eeeg-dataset>). Data were recorded at 128 Hz using an Emotiv Epoc Flex gel-based headset following the international 10-20 electrode placement system. Four experimental conditions were recorded per subject: (1) relaxation (75 s, used as personal baseline), (2) Stroop color-word interference task (75 s), (3) mental arithmetic task (75 s), and (4) mirror image recognition task (75 s). The three non-relaxation conditions constitute the stress conditions.

A data quality note: inspection of the raw recordings revealed inter-subject variability in signal quality, with some participants showing higher baseline noise floors and artifact rates than others. This heterogeneity is an inherent characteristic of the dataset and is one reason per-subject personalisation is methodologically preferable to pooled population models. Limitations include mild cognitive stressors without cortisol sampling, heart rate variability measurement, or self-report stress ratings to validate that participants experienced genuine physiological stress.

The 75-second trial duration also limits the number of baseline epochs that are available per subject. Additionally, epochs inherit labels from their parent task condition and not independently alternating stress states meaning that the temporal autocorrelation within trials may partially contribute to classification performance.

3.2 Preprocessing

All preprocessing was performed using MNE-Python 1.6 [14]. The following steps were applied in sequence to all 40 subjects across all four conditions.

Bandpass filtering (1-40 Hz): We applied a 4th-order zero-phase Butterworth bandpass filter from 1 Hz to 40 Hz. The lower cutoff of 1 Hz removes slow baseline drift while preserving delta-band neural activity. The upper cutoff of 40 Hz was chosen because the SAM40 data were recorded in Europe, where mains electrical interference occurs at 50 Hz; filtering below this frequency retains all neural signals of interest while avoiding the powerline artifact. A separate 50 Hz notch filter was additionally applied to suppress any residual powerline interference.

Average reference: We applied a common average reference (CAR), and removed the mean signal across all 32 channels from each channel at every time point. CAR is preferred over a single reference electrode when the reference electrode itself may carry neural signal, as is common in scalp EEG. It reduces the influence of any single electrode's noise on the entire recording and is standard practice in EEG affective computing research [3].

ICA artifact removal: Independent Component Analysis (FastICA, 15 components, random state 42, maximum 800 iterations) was applied to decompose the multichannel signal into statistically independent sources. Ocular artifact components were automatically identified using Fp1 as an electrooculogram (EOG) proxy channel with a z-score threshold of 2.5 standard deviations, and removed before back-projection. This step reduced peak-to-peak amplitude from $\pm 1,096$ μV to ± 181 μV , confirming successful removal of eye blink and saccade artifacts.

Wearable simulation and upper bound caveat: For the wearable condition, electrodes T7 and T8 were extracted after the full 32-channel preprocessing pipeline described above. This is an important methodological limitation: both common average reference and ICA decomposition exploit information from all 32 channels before T7/T8 extraction. A real two-electrode wearable device cannot perform 32-channel ICA or compute a 32-channel average reference. Current results therefore, represent an upper bound on achievable wearable performance. A wearable-feasible preprocessing pipeline using bipolar T7-T8 referencing and single-channel threshold-based artifact rejection is planned for Layer 2.

3.3 Feature Extraction: Differential Entropy

We used differential entropy (DE) as the spectral feature. The rationale for this choice over raw band power is that raw band power estimates are sensitive to absolute amplitude, which

varies substantially across subjects due to differences in skull thickness, electrode impedance, and individual neurophysiology. This makes it more stable across subjects, as DE captures the spectral shape instead of the absolute amplitude.

For a Gaussian-distributed signal with variance σ^2 , the differential entropy is given by $DE = (1/2) \log(2\pi e \sigma^2)$. Following Shi et al. [12] and established practice [3], for fixed-length EEG epochs this simplifies to the computationally efficient approximation $DE = \log(\text{var}(x_{\text{band}}) + \epsilon)$, where x_{band} is the bandpass-filtered epoch and ϵ is a small constant to avoid $\log(0)$.

DE was computed per channel, per frequency band, per 4-second sliding window with 50% overlap. Five standard frequency bands were used: delta (1-3 Hz), theta (4-7 Hz), alpha (8-13 Hz), beta (14-30 Hz), and gamma (31-40 Hz). For the wearable condition, only T7 and T8 were used, producing a 10-dimensional feature vector (2 channels x 5 bands). For the full-channel condition, all 32 channels produced a 160-dimensional feature vector.

3.4 Models

Personal Baseline Deviation Model (unsupervised, subject-specific): No stress labels are required. For each subject, the mean and standard deviation of each feature dimension are computed from relaxation-condition epochs only. Each following epoch is assessed by the L2 norm of its z-scored deviation from the individual's defined baseline distribution. Epochs exceeding the 90th percentile of the subject's own baseline deviation scores are classified as stress. The 90th-percentile threshold defines the top 10% of baseline deviation as aberrant, a conservative, specificity-first criterion following anomaly detection conventions for physiological signals. Sensitivity to this threshold choice is acknowledged as a limitation.

Population Classifier (supervised, cross-subject): A gradient boosting classifier (100 estimators, depth 4, learning rate 0.1) was trained on labelled data from all subjects except the held-out test subject, with feature standardisation per fold. Evaluation used leave-one-subject-out cross-validation (LOSO-CV). Classes were balanced by random undersampling within each fold.

Comparison framing: these two models differ in supervision requirements. The comparison is intentionally asymmetric: we ask whether zero-label personalisation at test time can match or exceed what a fully supervised cross-subject model achieves. This is the operationally relevant question for consumer wearable deployment, where collecting per-user labelled stress data is impractical.

3.5 Statistical Analysis

The Wilcoxon signed-rank test was used to compare per-subject accuracy distributions between the two models ($n=40$, two-tailed, $\alpha=0.05$). This nonparametric test is appropriate for paired, non-normally distributed accuracy data. Effect size was computed as $r = |z|/\sqrt{n}$.

Additional metrics for the population classifier include balanced accuracy, AUC-ROC, and F1 score (per LOSO fold, then averaged); median and interquartile range are reported for the baseline model. SHAP analysis [15] was applied to the population classifier only using TreeExplainer on the final LOSO fold.

4. Results

4.1 Model Comparison

Condition	Model	Accuracy	Bal. Acc.	AUC-ROC	F1	p	r
Wearable (T7+T8)	Personal baseline	0.611 +/- 0.134	--	--	--	0.025 *	0.35 5
	Population (LOSO)	0.538 +/- 0.069	0.538 +/- 0.069	0.562 +/- 0.117	0.551 +/- 0.121		
Full 32ch	Personal baseline	0.693 +/- 0.164	--	--	--	0.044 *	0.31 8
	Population (LOSO)	0.619 +/- 0.120	0.619 +/- 0.120	0.707 +/- 0.160	0.629 +/- 0.149		
Chance	--	0.500	0.500	0.500	--	--	--

Table 1. Main result. * $p < 0.05$ Wilcoxon signed-rank test. Blank metric cells indicate measures that are not applicable to the personal baseline deviation model, which produces unsupervised deviation scores rather than calibrated probabilistic outputs. Median (IQR): wearable baseline 0.556 (0.500-0.712); full 32ch baseline 0.674 (0.535-0.840).

The personal baseline model significantly outperformed the population classifier in both electrode configurations ($p < 0.05$, medium effect size). The personalisation advantage is 7.3 percentage points on wearable and 7.4 on full-channel, essentially identical, confirming that the gain is robust to electrode reduction. The wearable population classifier AUC of 0.562 is barely above chance, while the full 32-channel AUC of 0.707 is moderate. This contrast demonstrates that cross-subject discriminative patterns are substantially weaker in the temporal-only configuration, explaining why personalisation provides the larger relative benefit there. 18/40 subjects (45%) achieved baseline model accuracy ≥ 0.60 on the wearable condition; 28/40 (70%) on the full 32-channel condition.

Personalized baseline vs population classifier for EEG stress detection
Wearable simulation (T7+T8) vs full 32-channel EEG -- SAM40 dataset, 40 subjects

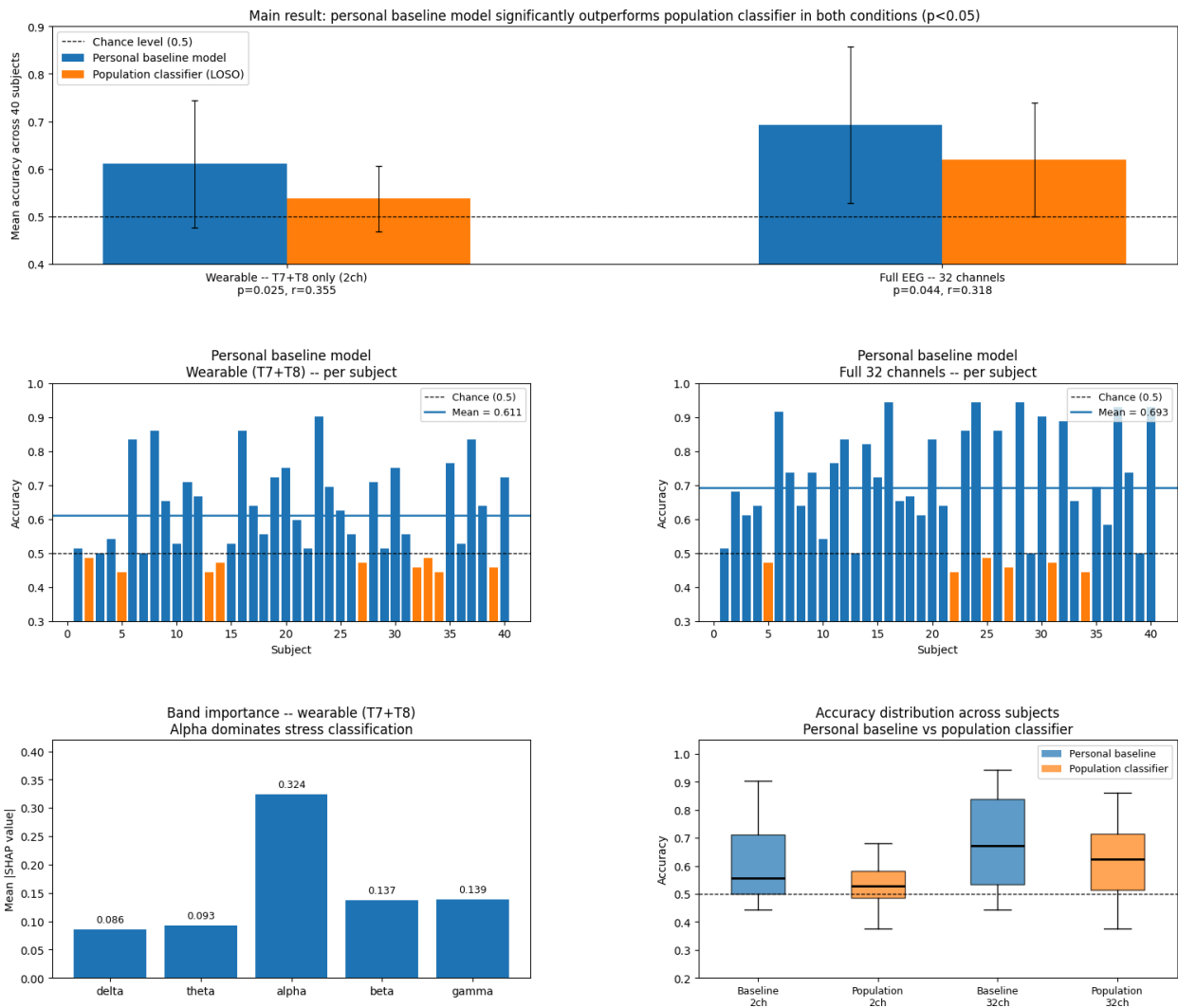


Figure 1. Main model comparison results across wearable (T7/T8) and full 32-channel conditions. Top: mean classification accuracy for personal baseline and population LOSO models with Wilcoxon significance testing. Middle: per-subject accuracy distributions illustrating substantial inter-subject variability. Bottom left: SHAP feature importance for the wearable population classifier showing dominant temporal alpha contribution. Bottom right: boxplot distributions across all 40 subjects. Personalisation significantly improves performance under both electrode configurations.

4.2 SHAP Feature Attribution (Population Classifier)

SHAP analysis was applied to the population classifier on the final LOSO fold. These results characterise what discriminates stress from relaxation across subjects in a supervised learning context and do not explain the personalisation mechanism of the baseline deviation model, for which SHAP is not applicable. For the wearable condition, temporal alpha DE dominates (mean $|\text{SHAP}| = 0.324$), 2.4x greater than beta (0.137). This finding independently recovers alpha suppression as the primary stress biomarker without any alpha-specific feature engineering, consistent with the neuroscience literature [4, 6]. For full 32 channels, frontal features dominate: Fp2 alpha (0.316), CZ theta (0.277), Fp1 beta (0.227). The performance gap between conditions is directly attributable to the unavailability of frontal channels on a temporal-only device.

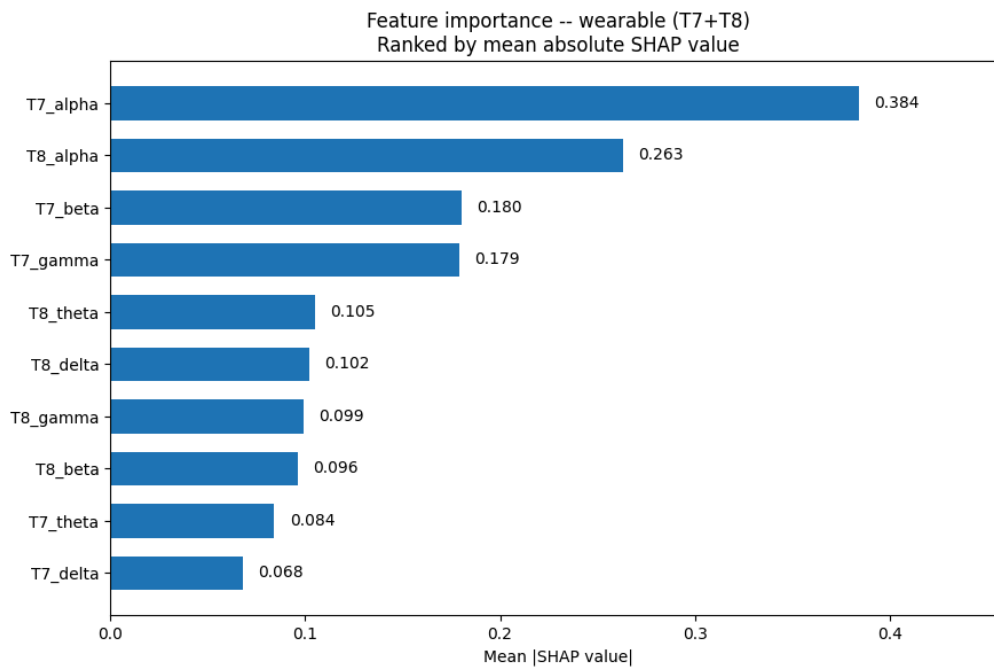


Figure 2. SHAP feature importance for the wearable population classifier (T7/T8 configuration). Temporal alpha differential entropy contributes most strongly to cross-subject stress discrimination that exceeds beta-band contribution by approximately 2.4x.

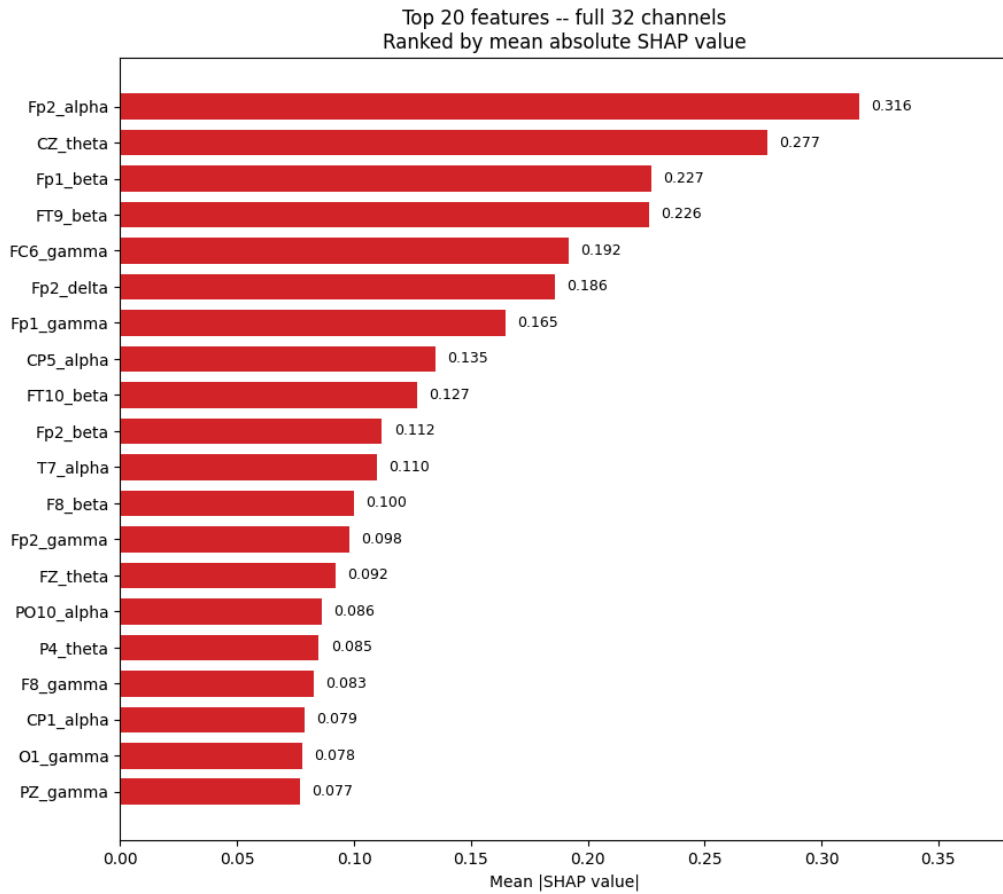


Figure 3. SHAP feature importance for the full 32-channel population classifier. Frontal alpha and beta features dominate classification importance and explain the performance gap between temporal-only wearable simulation and full-scalp EEG analysis.

4.3 Extended Analyses

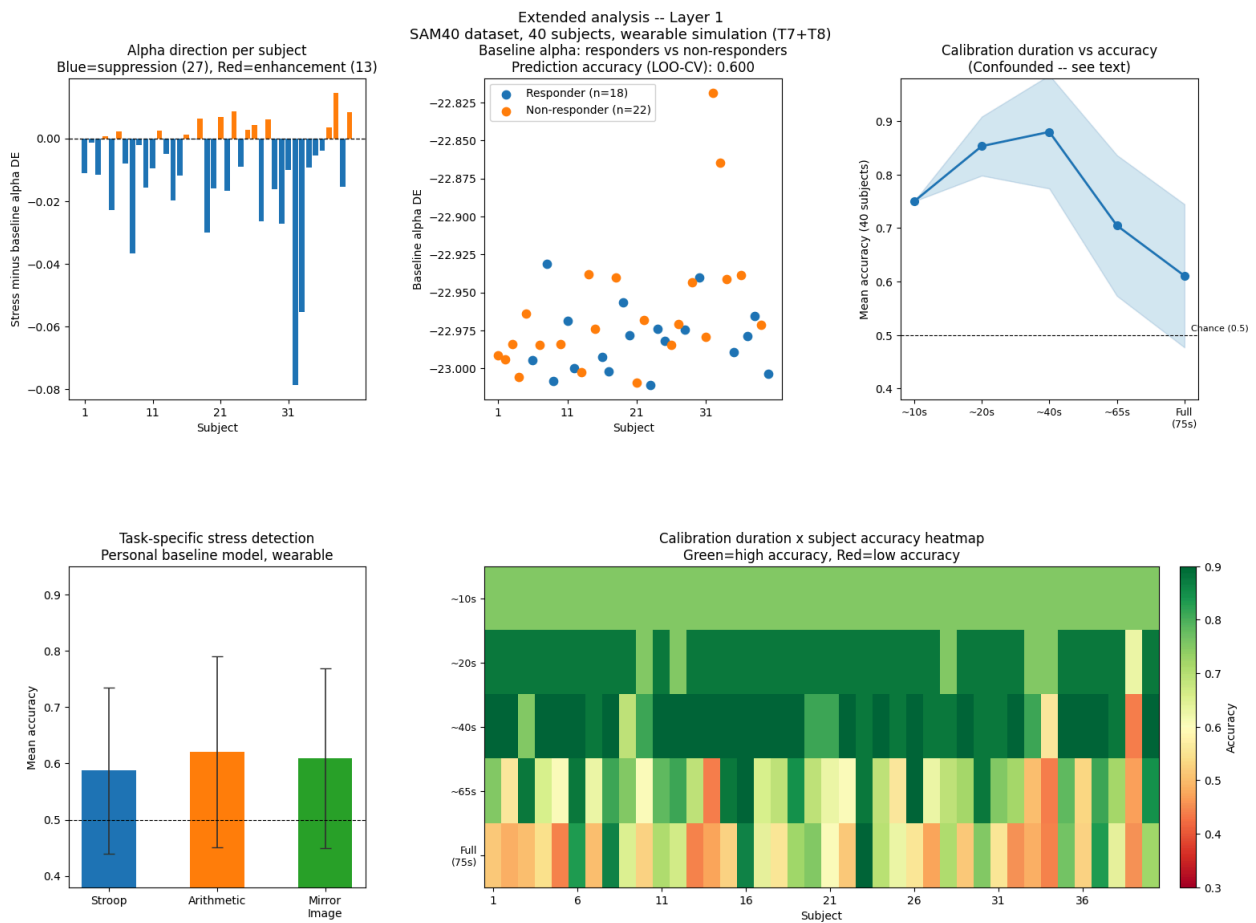


Figure 4. Extended analyses exploring heterogeneity and calibration effects. Top left: direction of alpha change under stress by subject. Top center: baseline alpha differential entropy for responders versus non-responders. Top right: calibration duration analysis (confounded within-session estimate; interpreted qualitatively only). Bottom left: task-specific classification performance. Bottom right: subject-by-calibration heatmap that illustrates strong inter-subject variability.

Alpha direction (Analysis 1): 27/40 subjects (67.5%) show alpha suppression under stress; 13/40 (32.5%) show enhancement. Suppression direction significantly predicts responder status ($r=0.338$, $p=0.033$). Note: the SAM40 relaxation protocol does not specify eyes-open vs eyes-closed conditions, which may confound alpha comparisons; this is flagged as a limitation.

Responder prediction (Analysis 2): Logistic regression from four baseline EEG characteristics achieves LOO-CV accuracy of 0.600 vs 0.550 chance. Signal stability is the strongest predictor (coefficient +0.824). Results are exploratory at $n=40$.

Calibration duration (Analysis 3): This analysis is confounded because baseline and test data were drawn from the same 75-second recording. Shorter calibration windows artificially inflate accuracy. The full 75-second result (0.611) is the only ecologically valid estimate. This is reported as a qualitative pilot only.

Task-specific accuracy (Analysis 4): All three tasks produce above-chance classification (all $p < 0.01$). Arithmetic produces the highest accuracy (0.620 +/- 0.169), Stroop the lowest (0.587 +/- 0.148), consistent with arithmetic engaging autonomic stress pathways more reliably than attentional interference.

5. Discussion

Four findings carry practical relevance for consumer wearable EEG design. Personalisation advantage persists under electrode reduction. The 7.3-percentage-point accuracy gain of the baseline model is nearly identical under wearable and full-channel conditions, which suggests that the benefit does not depend on spatial richness across the scalp. The wearable population classifier AUC of 0.562 is barely above chance, meaning that cross-subject stress patterns carry very little discriminative information in the temporal-only configuration. For 2-channel devices, this result supports investing in short per-user calibration workflows over more complex cross-subject models. Additionally, due to SAM40 lacking physiological validation measures like cortisol or heart-rate variability, the present findings are more precisely interpreted as task-deviation detection under cognitive-affective load.

Temporal alpha DE is the primary wearable stress biomarker. SHAP attribution independently recovers alpha suppression as the dominant cross-subject feature at T7/T8, consistent with the established neuroscience literature [4, 6]. The performance gap between wearable and full-channel conditions is attributable to the loss of frontal access; a device adding one frontal electrode such as AF7 or AF8 would likely recover a substantial portion of this gap.

32.5% of subjects show alpha enhancement rather than suppression, meaning that population classifiers with fixed directional assumptions systematically misclassify this subgroup. Therefore, tracking signed rather than magnitude deviation in the alpha band is a key target improvement for Layer 2.

Baseline signal stability predicts responder status, as subjects with more stable resting EEG are more likely to be correctly classified, which in turn suggests that a brief resting-state recording at device startup could serve as a calibration quality screen, analogous to fingerprint quality checks in biometric authentication.

More broadly, the present findings support the idea that wearable EEG deployment may benefit more from lightweight longitudinal adaptation than from increasingly complex static population models. Future consumer neurotechnology systems may therefore require continual subject-specific recalibration to maintain robustness under nonstationary real-world conditions.

5.1 Study Limitations

The following limitations are explicitly acknowledged. They define the boundaries within which results should be interpreted and do not invalidate the findings.

- Insufficient data: $n=40$ subjects with 75-second trials and mild cognitive stressors. No single dataset is ever large enough to satisfy all generalisability requirements; these findings are best understood as a pilot warranting replication with stronger stress induction and more subjects.
- Incomplete experimental coverage: Many validation tests (ROC threshold optimisation, alternative unsupervised detectors such as one-class SVM or isolation forest, domain adaptation baselines) were not performed. These are targets for Layer 2.
- Wearable simulation upper bound: The wearable condition benefits from preprocessing operations (32-channel CAR and ICA) unavailable to true two-electrode consumer systems. Therefore, current results should be interpreted as an upper-bound estimate of wearable-feasible performance rather than direct real-world device accuracy.
- No stress validation: SAM40 contains no cortisol, heart rate variability, or self-report data confirming genuine physiological stress.
- Eyes-open/closed unspecified: The SAM40 relaxation protocol does not specify visual fixation or eyes-closed conditions, potentially confounding alpha direction analysis.
- SHAP on single fold: SHAP was computed on the final LOSO fold only; variance across folds is not reported.
- Calibration duration confounded: within-session overlap inflates short-window calibration results.
- Single classifier baseline: The study compares against a single interpretable cross-subject baseline classifier. More advanced approaches including Riemannian geometry features, domain adaptation, and covariance-alignment methods were not evaluated and constitute a primary target for Layer 2 comparative benchmarking.

6. Conclusion

This pilot study demonstrates that a simple and unsupervised personalisation approach significantly outperforms a supervised cross-subject population classifier for EEG stress detection, even when analysis is restricted to two temporal electrodes simulating a consumer wearable device. The personalisation advantage is statistically significant with a medium effect size and is robust to electrode reduction. SHAP analysis identifies temporal alpha differential entropy as the primary wearable stress biomarker, and the existence of a consistent alpha-enhancing subpopulation (32.5% of subjects) highlights the importance of directional personalisation over fixed-direction assumptions. These findings are modest in absolute accuracy terms and rest on a mild-stressor dataset without physiological stress validation; they are best interpreted as strong motivation for the original data collection planned in Layer 2.

Code and preprocessing pipelines are not publicly released at this stage due to ongoing longitudinal development and planned Layer 2 validation experiments. The SAM40 dataset is available at <https://www.kaggle.com/datasets/benbenbite/sam40-eeeg-dataset>. All random seeds are fixed for exact reproducibility.

Ethics statement. This study is a computational analysis of a publicly available, fully anonymised dataset. No new human participants were recruited. The SAM40 dataset was collected under appropriate ethical oversight by its original authors; all data used here are secondary, anonymised, and freely available for research purposes.

7. Future Directions

My goal is to carry this work from computational analysis into hands-on laboratory implementation. Layer 2 will collect original EEG data using a clinical 32-channel system with TSST stress induction [13]: five minutes of speech preparation, five minutes of performance before neutral judges, and five minutes of mental arithmetic, with five minutes of resting baseline before and after. The TSST reliably activates the hypothalamic-pituitary-adrenal (HPA) axis and produces cortisol elevation in the majority of healthy adults [13], providing the physiological ground truth absent in SAM40. A minimum of 15 participants will be recruited. Four methodological improvements will be implemented: (1) a wearable-feasible preprocessing pipeline using bipolar T7-T8 referencing and single-channel artifact rejection, to bound the simulation-to-deployment gap quantified here; (2) separate recording sessions for baseline calibration and TSST evaluation, eliminating the within-session calibration confound; (3) a directional deviation score tracking signed alpha changes will be compared against the magnitude-based score used here; and (4) State-Trait Anxiety Inventory self-report ratings at four timepoints will provide subjective stress ground truth.

Layer 3 will test the key findings on real consumer wearable hardware to validate the simulation assumption underlying Layer 1 and quantify the real-world performance gap. I aim to learn and implement this full research pipeline in person within a laboratory setting, and thereafter to use the experience to design and execute further investigations in neurotech stress monitoring.

References

- [1] Saha S and Baumert M (2020) Intra- and Inter-subject Variability in EEG-Based Sensorimotor Brain Computer Interface: A Review. *Front. Comput. Neurosci.* 13:87. doi: 10.3389/fncom.2019.00087
- [2] Apicella, Andrea, et al. "Toward cross-subject and cross-session generalization in EEG-based emotion recognition: Systematic review, taxonomy, and methods." *Neurocomputing* 604 (2024): <https://doi.org/10.1016/j.neucom.2024.128354>.
- [3] Fdez J, Guttenberg N, Witkowski O and Pasquali A (2021) Cross-Subject EEG-Based Emotion Recognition Through Neural Networks With Stratified Normalization. *Front. Neurosci.* 15:626277. doi: 10.3389/fnins.2021.626277
- [4] Schapkin SA, Raggatz J, Hillmert M, Böckelmann I. EEG correlates of cognitive load in a multiple choice reaction task. *Acta Neurobiol Exp (Wars)*. 2020;80(1):76-89. PMID: 32214277.
- [5] Vos G, Ebrahimpour M, van Eijk L, Sarnyai Z, Rahimi Azghadi M. Stress monitoring using low-cost electroencephalogram devices: A systematic literature review. *Int J Med Inform.* 2025 Jun;198:105859. doi: 10.1016/j.ijmedinf.2025.105859. Epub 2025 Mar 6. PMID: 40056845.
- [6] Berretz G, Packheiser J, Wolf OT, Ocklenburg S. Acute stress increases left hemispheric activity measured via changes in frontal alpha asymmetries. *iScience*. 2022 Feb 1;25(2):103841. doi: 10.1016/j.isci.2022.103841. PMID: 35198894; PMCID: PMC8850739.
- [7] Athavipach C, Pan-Ngum S, Israsena P. A Wearable In-Ear EEG Device for Emotion Monitoring. *Sensors (Basel)*. 2019 Sep 17;19(18):4014. doi: 10.3390/s19184014. PMID: 31533329; PMCID: PMC6767669.
- [8] Moumane H, Pazuelo J, Nassar M, Juez JY, Valderrama M and Le Van Quyen M (2024) Signal quality evaluation of an in-ear EEG device in comparison to a conventional cap system. *Front. Neurosci.* 18:1441897. doi: 10.3389/fnins.2024.1441897
- [9] Koelstra, Sander & Mühl, Christian & Soleymani, Mohammad & Lee, Jong-Seok & Yazdani, Ashkan & Ebrahimi, Touradj & Pun, Thierry & Nijholt, Anton & Patras, Ioannis. (2011). DEAP: A Database for Emotion Analysis Using Physiological Signals. *IEEE Transactions on Affective Computing*. 3. 18-31. 10.1109/T-AFFC.2011.15.
- [10] Lebepe, F. & Niezen, G. & Hancke, G. & Ramotsoela, Daniel. (2016). Wearable stress monitoring system using multiple sensors. 895-898. 10.1109/INDIN.2016.7819288.
- [11] Ju, Xiangyu & Li, Ming & Tian, Wenli & Hu, Dewen. (2023). EEG-based emotion recognition using a temporal-difference minimizing neural network. *Cognitive Neurodynamics*. 18. 10.1007/s11571-023-10004-w.
- [12] Kirschbaum C, Pirke KM, Hellhammer DH. The 'Trier Social Stress Test'--a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology*. 1993;28(1-2):76-81. doi: 10.1159/000119004. PMID: 8255414.
- [13] Gramfort A, Luessi M, Larson E, Engemann DA, Strohmeier D, Brodbeck C, Goj R, Jas M, Brooks T, Parkkonen L and Hämäläinen M (2013) MEG and EEG data analysis with MNE-Python. *Front. Neuroinform.* 7:267. doi: 10.3389/fnins.2013.00267

[14] Lundberg, Scott & Lee, Su-In. (2017). A Unified Approach to Interpreting Model Predictions.
10.48550/arXiv.1705.07874.