Case 2

02582 Computational Data Analysis

s260170, s260219, s257659

May 5, 2026

Introduction

The *EmoPairCompete* dataset contains biosignals (heart rate, electrodermal activity, and skin temperature) from 26 participants performing a stressful tangram puzzle competition in teams. Each participant was recorded across four rounds, with each round comprising a pre-task rest, a competitive puzzle phase, and a post-task rest, yielding twelve recordings per participant. From these recordings, 51 summary features were extracted, alongside self-reported emotion ratings collected after each phase.

The central interest is whether physiological state during stressful task engagement can be distinguished from physiological state at rest. This is complicated by large inter-subject baseline differences: some people sweat more, run warmer, or have higher resting heart rates than others, and these differences can dwarf the experimental effect of interest.

We approach this as an exploratory representation-learning problem and address the following research questions:

1. How is variance in the biosignal feature space distributed between inter-subject differences and experimental phase?
2. Do linear representations (PCA) make experimental phase separable, and does separability improve when individual baselines are removed?
3. To what extent are biosignal features and self-reported emotion ratings linearly related across modalities?
4. Do non-linear methods (t-SNE, UMAP) reveal structure that linear methods miss?

Data and preprocessing

Dataset

The dataset comprises 312 recordings (26 individuals $\times$ 4 rounds $\times$ 3 phases). Each row corresponds to one phase of one round for one participant, with 51 numerical biosignal features (heart rate, EDA tonic,

EDA phasic, temperature; including means, medians, standard deviations, extrema, AUC, kurtosis, skewness, slope statistics, and EDA peak counts), 11 self-reported emotion ratings on the I-PANAS-SF scale, and metadata identifying the individual, round, phase, cohort, and team. The data is perfectly balanced. Crucially, observations are not independent: each individual contributes 12 recordings, so naive random splits would leak subject-specific information into evaluation. We therefore treat the individual as the unit of generalization throughout.

Only 9 of approximately 21,840 cells were missing: seven emotion ratings (PANAS items skipped by participants, four rows affected) and two EDA-phasic timing features for a single row where a peak was detected but its rise and recovery times could not be characterized (likely truncated at the recording boundary). We dropped this single row for biosignal-only analyses (311 rows) and additionally dropped the four emotion-NaN rows for cross-modal analyses (307 rows).

Standardization

Biosignal features span roughly nine orders of magnitude in scale (from $\sim 3 \times 10^{-5}$ for temperature slopes up to $\sim 4 \times 10^{4}$ for temperature AUC), so standardization is mandatory before any variance- or distance-based method. The choice of which data the standardization parameters are computed from has substantial consequences for what subsequent methods recover, so we compare three pipelines:

- **D1 (global):** mean and standard deviation computed across all 311 rows. Retains both inter- and intra-subject variation.
- **D2 (per-subject):** mean and standard deviation computed within each individual. Removes inter-subject variation by construction.
- **D3 (baseline-relative):** for each individual, statistics computed across that individual's resting-phase rows only; all 12 rows are then re-expressed as deviation from this resting baseline.

This comparison directly addresses Research Question 2. We use StandardScaler (mean/standard deviation) rather than RobustScaler; a brief diagnostic showed that the IQR-based scaling counter-intuitively amplified outlier influence on the leading components, because EDA-phasic features have very tight bulk distributions with rare large spikes that become more extreme after IQR rescaling. Specifically, features like *EDA_TD_P_Min* exhibit an extreme skewness of approximately -8.8 and an excess kurtosis of over 100, which causes the IQR-based approach to fail.

Outliers

Diagnostic investigation revealed that one individual (Individual 19) accounted for 9 of the 10 most extreme PC1 scores under D1, with extremeness concentrated entirely in EDA features (medians 5–14 standard deviations above the population) and normal heart rate and temperature. This is consistent with the well-documented phenomenon of large inter-individual differences in electrodermal reactivity, and we treated it as genuine physiological variation rather than measurement artifact. No rows were excluded; the dominance of inter-subject extremes under D1 motivates the per-subject pipelines D2 and D3.

Methods

We applied three representation-learning methods to the biosignal feature space and one cross-modal method linking biosignals to self-reported emotions, with a unified validation strategy described at the end of this section. All analyses used scikit-learn (PCA, t-SNE, CCA, LogisticRegression) and the umap-learn package (UMAP). Each method was run independently on each of the three preprocessing pipelines (D1, D2, D3) so that the effect of preprocessing could be assessed separately from the effect of method choice.

Principal Component Analysis (PCA)

PCA finds an orthogonal rotation of the feature space such that the first axis (PC1) maximizes variance, the second maximizes residual variance, and so on. We retained the first five components for downstream evaluation, and report explained variance ratios (see Figure 1), loadings to interpret what each component represents physiologically, and scatter plots of PC1 against PC2 colored by phase and by individual. Because PCA is unsupervised, any alignment between principal components and experimental phase is a finding, not a constraint.

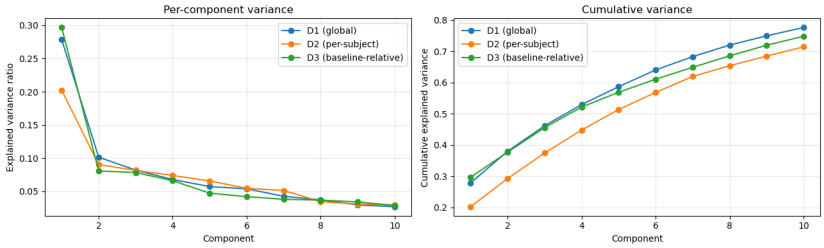

Figure 1: PCA explained variance ratios (per-component, left; cumulative, right) under each preprocessing pipeline. D2 flattens the variance curve because per-subject standardization removes the largest single source of variance.

Canonical Correlation Analysis (CCA)

CCA finds pairs of directions, one in each of two views, such that the projections of the data onto those directions are maximally correlated across observations. We applied CCA with biosignal features as one view ($X \in \mathbb{R}^{307\times51}$) and emotion ratings as the other ($Y \in \mathbb{R}^{307\times11}$), both standardized. Up to five canonical pairs were extracted, with detailed analysis focused on the first three. CCA is well known to overfit when feature dimensionality is large relative to sample size, so we paid particular attention to validation (below).

Non-linear embeddings

We applied t-SNE (perplexity = 30, PCA initialization, random state = 42) and UMAP (`n_neighbors` = 15, `min_dist` = 0.1, random state = 42), both producing 2D embeddings. Unlike PCA, these methods preserve local neighborhoods rather than global variance, and their axes carry no direct interpretation. We used them as a comparison: if non-linear methods reveal phase structure invisible to PCA, the underlying signal is non-linear; otherwise PCA captures whatever structure is present.

Validation strategy

The repeated-measures structure of the dataset (12 recordings per individual) makes random train/test splits invalid, since subject-specific baselines would leak across folds. We used **Leave-One-Subject-Out cross-validation** (LOSO-CV) throughout: for each of the 26 individuals, the model is trained on the remaining 25 individuals' recordings and evaluated on the held-out individual. This estimates how well the learned representation generalizes to physiologies the model has never seen.

For PCA and the non-linear embeddings, we evaluated representation quality by training a multinomial logistic regression on the embedded features to predict phase, reporting LOSO-CV balanced accuracy (chance = 33.3%). For CCA, we evaluated the canonical correlations themselves under LOSO: for each held-out subject, CCA was fit on the remaining 25 subjects, the held-out subject was projected into canonical space, and the canonical correlations were computed on the pooled out-of-sample projections. The gap between in-sample and LOSO canonical correlations directly quantifies overfitting.

Results

Variance Partitioning and Linear Separability

The initial variance decomposition revealed that a substantial portion of the biosignal variance is driven by inter-subject differences rather than the experimental phases. Under global standardization (D1), 45.1% of total feature-space variance was between-subject, compared to only 2.4% between-phase — an 18-fold dominance of inter-individual differences over experimental condition. A logistic regression model on the first five PCs of D1, evaluated via LOSO-CV, achieved a balanced accuracy of 48.0% (chance = 33.3%). This subject dominance is explicitly visualized in Figure 2, where observations cluster tightly by individual physiology. The experimental phase explained a negligible amount of variance in the leading principal components under D1 (1.16% for PC1 and 3.61% for PC2).

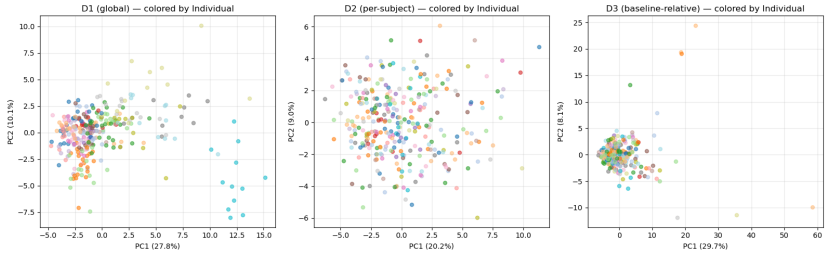

Figure 2: Scatter plot of PC1 vs PC2 under global standardization (D1), colored by individual. The local groupings demonstrate that inter-subject variability is the primary driver of variance in the raw biosignal space.

Removing subject-specific baselines significantly improved the linear representation of the experimental task. Under per-subject (D2) and baseline-relative (D3) standardization, LOSO phase-prediction balanced accuracy increased to 57.4% and 56.1%, respectively. The phase effect on the principal components grew substantially: under D2, PC1 and PC2 reached 11.8% and 19.9% phase η^2, while under D3, PC1 alone reached 18.3%.

An analysis of the loadings for D3 reveals that **PC1 is heavily dominated by phasic electrodermal activity (EDA)**, specifically the median of the phasic EDA (*EDA_TD_P_Median*, loading = 0.605) and

the area under the curve (*EDA_TD_P_AUC*, loading = 0.382). In contrast, PC2 primarily captures tonic EDA slopes and phasic minimums in the baseline-relative pipeline. The emergence of phase separability after D3 normalization is illustrated in Figure 3.

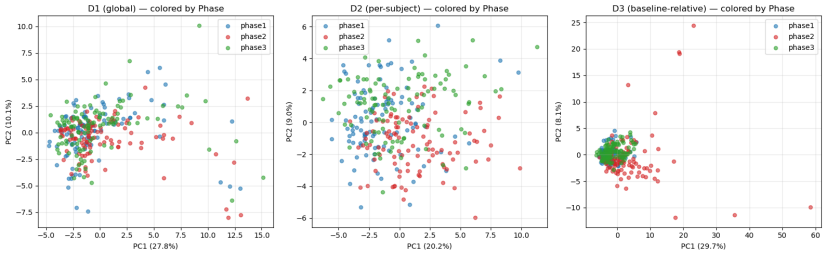

Figure 3: Scatter plot of PC1 vs PC2 under baseline-relative standardization (D3), colored by experimental phase (rest vs puzzle). Adjusting for resting baselines effectively reveals the linear separation between states.

Cross-Modal Relationships and Generalization

We applied CCA with biosignal features and emotion ratings as the two views, extracting up to five canonical pairs under each preprocessing pipeline. The first canonical correlations in-sample appeared strong ($\rho_1 = 0.67$ for D1 and D3, $\rho_1 = 0.60$ for D2) but collapsed dramatically under LOSO cross-validation (Table 1). Beyond the first pair, ρ_2 and ρ_3 collapsed to essentially zero ($|\rho| < 0.05$) across all pipelines, indicating that no meaningful second or third cross-modal direction exists in this dataset.

Pipeline	ρ_1 in-sample	ρ_1 LOSO	ρ_2, ρ_3 LOSO
D1 (global)	0.67	0.29	~ 0
D2 (per-subject)	0.60	0.32	~ 0
D3 (baseline-relative)	0.67	0.04	~ 0

Table 1: In-sample vs LOSO canonical correlations. The gap quantifies overfitting; D3's near-total collapse indicates that all of its in-sample structure was subject-specific.

The pipeline ordering inverts compared to PCA: D2 yielded the most generalizable cross-modal structure ($\rho_1 = 0.32$), while D3 — the best pipeline for unimodal phase separability — collapsed almost entirely ($\rho_1 = 0.04$). This asymmetry is informative: D3's per-subject baseline subtraction removes precisely the between-subject variance that CCA exploits to find shared cross-modal structure across the population. PCA benefits from removing inter-subject variation because phase signal then becomes proportionally larger; CCA suffers from removing it because the cross-modal correlations in this dataset are themselves predominantly between-subject in nature.

The first canonical direction under D2 links a biosignal pattern of high phasic EDA AUC (0.420), high temperature median (0.373), and low tonic EDA median (-0.405) to an emotion cluster of high hostility (0.589), frustration (0.505), and active state (0.440). Physiologically, this corresponds to a sympathetic-arousal pattern (increased phasic skin conductance responses) aligning with a high-activation negative-affect pattern in self-report — a coherent and interpretable axis, though modest in magnitude.

Non-linear comparison

To address Research Question 4, we applied t-SNE and UMAP to each preprocessing pipeline and evaluated the resulting 2D embeddings using the same LOSO logistic regression protocol used for PCA (Table 2).

Pipeline	PCA (5 PCs)	t-SNE (2D)	UMAP (2D)
D1 (global)	48.0%	33.3%	35.6%
D2 (per-subject)	57.4%	47.9%	44.8%
D3 (baseline-relative)	56.1%	52.5%	47.3%

Table 2: LOSO-CV phase-prediction balanced accuracy (chance = 33.3%). Non-linear embeddings did not exceed PCA in any pipeline.

Two observations are noteworthy. First, the pipeline ranking (D1 $<$ D2 $\approx$ D3) is preserved across all methods, confirming that per-subject baseline correction is the dominant factor for revealing phase structure regardless of dimensionality reduction. Second, non-linear methods did not exceed PCA in any pipeline; under D1, t-SNE collapsed to chance level. Visual inspection of the embeddings (Figure 4) showed no clear phase clustering, while inter-subject groupings remained visible (notably Individual 19 forming a distinct sub-cluster in UMAP-D1). The interpretation is straightforward: the phase signal in this dataset is approximately linear, and applying non-linear embeddings on top does not extract additional information beyond what PCA captures. We note that PCA used 5 components against the non-linear methods' 2 dimensions, so part of PCA's advantage is dimensional; however, even the best non-linear pipeline (t-SNE-D3) did not exceed the corresponding PCA pipeline.

Discussion

Our results highlight three findings. First, inter-subject variability is the dominant source of variance in this dataset and must be removed before any phase signal becomes accessible: per-subject normalization roughly doubles the phase-related variance in the leading principal components and improves LOSO phase prediction from 48% to 57%. Second, classical CCA produces in-sample correlations ($\rho_1 \approx 0.67$) that overstate cross-modal structure by a factor of two or more under proper validation. The methodological implication is direct: in-sample canonical correlations on small biosignal datasets should be treated with skepticism. Third, the relative ranking of preprocessing pipelines depends on what is being asked: D3 (baseline-relative) is best for revealing within-subject phase structure in unimodal PCA, but worst for cross-modal CCA, because it removes precisely the between-subject variance that drives the generalizable cross-modal signal.

The phase signal we recover is real but modest: $\sim$57% balanced accuracy on a three-class problem under strict subject-level validation. The signal appears to be linear in nature — non-linear methods (t-SNE, UMAP) did not exceed PCA in any pipeline, suggesting there is no curved or manifold-shaped structure in the feature space for them to discover. Phasic EDA emerged as the dominant biosignal axis on which the experimental contrast manifests, consistent with its known role as an indicator of sympathetic activation in response to discrete stressors.

The CCA results indicate that the mapping between physiology and self-reported emotion is largely between-subject and weakly generalizable: people who report more frustration and hostility tend to show specific physiological patterns, but the within-subject changes induced by the puzzle do not produce a strong universal emotion-physiology coupling.

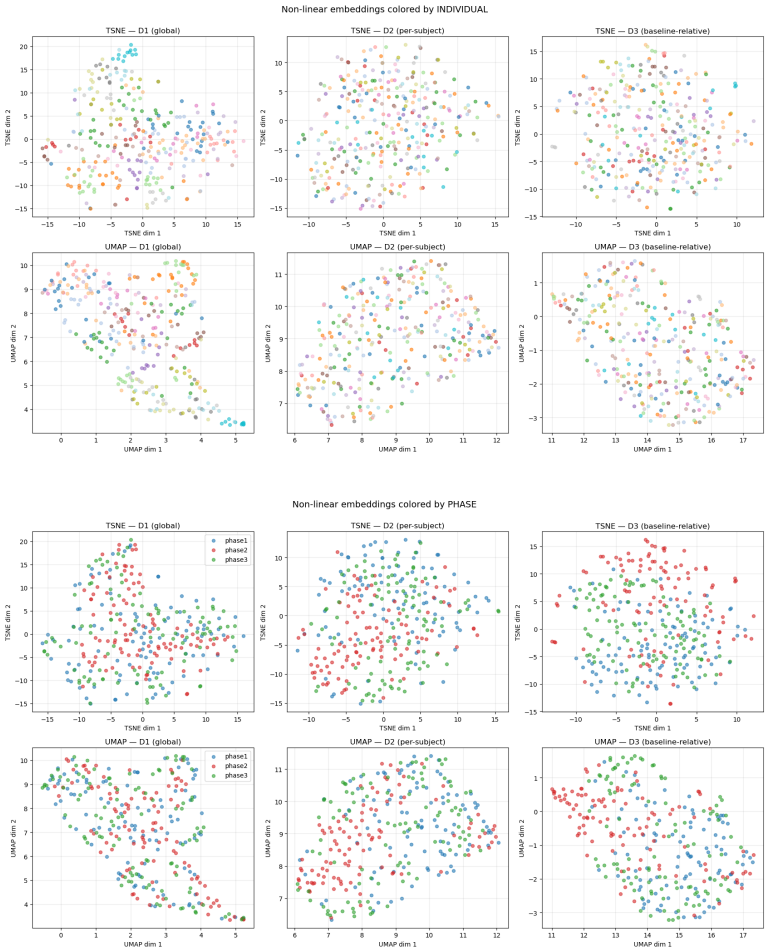

Figure 4: Non-linear embeddings (t-SNE, top half of each panel; UMAP, bottom half) under each preprocessing pipeline. **Top:** colored by individual, showing residual subject-level grouping under D1 (notably Individual 19's distinct sub-cluster in UMAP-D1). **Bottom:** colored by phase, showing no clear phase clustering across any pipeline.

Conclusion

We applied PCA, CCA, and non-linear embedding methods to physiological signals from a stress-induction experiment, comparing three preprocessing pipelines under Leave-One-Subject-Out validation. The dominant axis of variation in the data is between-subject differences, not the experimental condition. Removing individual baselines reveals a weak but real and generalizable phase signal ($\sim$57% LOSO accuracy vs. 33% chance) carried primarily by phasic electrodermal activity. Cross-modal structure between physiology and self-reported emotion is modest ($\rho_1 = 0.32$ under per-subject standardization) and predominantly between-subject in nature, with the in-sample canonical correlations overstating the true relationship by a factor of two or more. Non-linear methods did not extract additional information beyond PCA, suggesting the recovered signal is approximately linear. Future work could investigate temporal dynamics within recordings — which the present feature aggregation discards — and self-supervised representations trained on the raw biosignal time series.

Code Availability

All the source code for data preprocessing, visualization, and model validation is hosted on GitHub. The repository includes Jupyter notebooks for initial EDA, preprocessing decisions, PCA, and CCA combined with non-linear manifold learning (t-SNE/UMAP).
Repository URL: https://github.com/albertcasajuana01/cda_case2

Acknowledgement of GenAI use

In the preparation of this report, generative AI tools were used to assist with the structural editing, proofreading, and stylistic refinement of the text to ensure academic clarity. The core methodology, experimental design, data analysis, and critical interpretation of the results are entirely the original work of the authors.