Case 2

02582 Computational Data Analysis

s233498 & s233593 & s233494 & s236771

October 15, 2024

1 Introduction

1.1 Background

Physiological signals provide critical insights into emotional and stress responses, essential for applications in health monitoring and mental health treatment. The "EmoPairCompete" dataset, introduced by Das et al. (2024) [2], offers a unique perspective on emotion and frustration assessment under team and competitive behaviors [2]. This dataset not only captures physiological signals such as heart rate, electrodermal activity, and temperature but also correlates these signals with self-reported emotional states, providing a comprehensive resource for studying physiological responses in semi-controlled settings.

1.2 Motivation

Despite the availability of biosignal datasets, significant challenges persist in understanding and quantifying individual variability in physiological responses to controlled stimuli. Previous studies have often relied on data from highly controlled settings, which may not adequately represent the complex and varied reactions individuals exhibit to similar conditions. [2] This variability underscores the need for analytical methods, such as clustering, to identify and analyze patterns in physiological responses that could generalize across different individuals under equivalent experimental conditions.

1.3 Research Question

This project leverages the EmoPairCompete dataset to explore whether physiological responses, as grouped by clustering techniques, exhibit consistency across different participants when exposed to the same experimental conditions. Specifically, it asks: *"To what extent do participants with similar biosignals transition across the experimental phases together, suggesting a common physiological reaction to the experimental conditions?"*

Or in other words, the goal is to find whether participants who are in the same box (cluster) in experiment phase one, will also move to the same box under the conditions imposed by phase two or three. To answer this question, we will apply clustering analysis to the physiological data collected across the three experimental phases, assessing the consistency of these clusters across different participants. This analysis will help determine if there is a uniform pattern in how participants physiologically respond to the structured experimental challenges posed by the team-based puzzles and competition setup in the dataset.

2 Data

2.1 Description

The dataset was collected from an experiment where participants engaged in a competitive team puzzle task. The experiment had three phases: pre-puzzle, puzzle and post-puzzle and was repeated for four rounds. For each participant in the experiment, physiological signals were recorded, including heart rate, electrodermal activity, and temperature alongside self-reported emotional states.
The provided dataset consists of 67 features and 312 observations ($p = 67$, $N = 312$), derived from the EmoPairCompete dataset. Out of the 67 features, 51 represent summary statistics for

biosignals: HR, TEMP, EDA_P, and EDA_T. Each biosignal is described by 12 summary statistics, including AUC, kurtosis, max, mean, median, min, skew, slope, max slope, mean slope, min slope, and standard deviation. Additionally, EDA_T incorporates peaks, RT, and ReT in its summary statistics.

The data also contains the self-reported emotional states of participants, capturing frustration levels and 10 emotions. The frustration levels are rated on a scale of 0 to 10, while the 10 emotion features are rated on a scale of 0 to 5. The 10 emotions features are: upset, hostile, alert, ashamed, inspired, nervous, attentive, afraid, active, and determined.

Furthermore, five of the data attributes are categorical: round, phase, individual, puzzler, and cohort. The "round" feature takes four distinct values, as the experiment was repeated four times. For the sake of our analysis, it is important to note that rounds 1 and 3 were conducted in the morning, while rounds 2 and 4 took place in the evening. The "phase" feature has 3 categories, which shows us which of the three: pre-puzzle, puzzle or post-puzzle the record belongs to. The "puzzler" feature is binary, indicating whether an individual is a puzzler or not. There are a total of 26 individuals, with the "individual" feature ranging from 1 to 26. Lastly, the "cohort" feature consists of six categories: D1_1, D1_2, D1_3, D1_4, D1_5, and D1_6, representing the acquisition rounds for which the data were collected.

In summary, for each individual, there are data from four rounds, with each round consisting of three phases. Thus, for each individual, we have a total of 12 observations. [2]

Feature	**Missing Values**
EDA_TD_P_RT	1
EDA_TD_P_ReT	1
inspired	2
attentive	1
afraid	1
active	1
determined	2

Table 1: Missing Values

2.2 Preprocessing

Since physiological data is inherently noisy and subject to variations across different conditions and individuals [4], and since clustering algorithms are sensitive to the scale of the features [5], several steps needed to be taken to prepare the dataset. The steps included: handling missing values, standardizing the data, and selecting the most relevant features.

2.2.1 Missing Values and Standardization

The dataset contained missing values in seven features (see Table 1). Since the number of missing values was small and since we had only 26 participants, we decided to replace the missing values with the mean of the respective feature. This was performed only for the physiological attributes, as the self-reported emotion scores were not used for the clustering analysis.

Additionally, to ensure that the clustering algorithms were not biased by the scale of the features, that contained different units and ranges, we standardized the data by subtracting the mean and dividing by the standard deviation of each feature. This ensured that all considered features contributed equally in the clustering process.

2.2.2 Feature Selection

With only 26 participants in our dataset and a high number of attributes, we needed to reduce the number of features for each individual to mitigate the curse of dimensionality. [5] This was done, as the increase in dimensions can make the data sparse, diminishing the meaning of distance metrics which are crucial for clustering algorithms like K-means. This sparsity can lead to overfitting and poor generalization of clusters. For this reason, we took several steps to reduce the dimensions and select the most informative features.

Correlation and Redundancy in Features

Since the features were derived from the same physiological signals, they exhibited the problem of multicollinearity as shown in the correlation matrix in Figure 1. To address this issue, we decided to remove one feature from each pair of highly correlated features with a threshold set to 0.9. This resulted in 33 attributes selected.

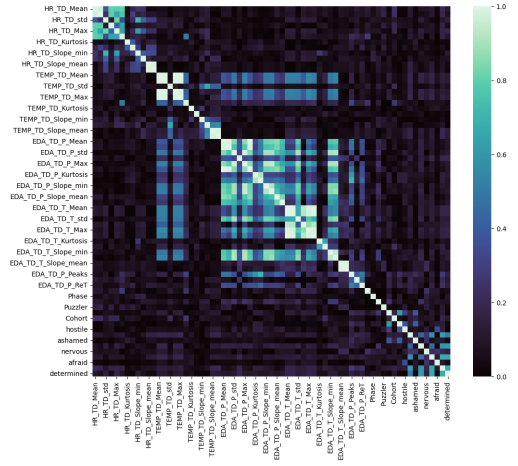


Figure 1: Absolute correlation matrix of pairwise combinations of features

Selecting Phase Sensitive Features

Following the reduction of correlated attributes, we examined how the remaining features varied across the experimental phases (pre-puzzle, puzzle, post-puzzle). Our focus was on identifying features that demonstrated significant changes with the experimental conditions, as these are more informative for answering our research question and understanding how the participants' bodies respond to the stress inducing activity and the subsequent calming down phase. For instance, while EDA_TD_Peaks showed considerable variation and were indicative of response to experimental conditions, features like mean temperature (TEMP_TD_mean) remained relatively constant and were less informative (see Figure 2).

By selecting features based on their sensitivity to phase changes, we ensured that our clustering was driven by physiological changes relevant to the experimental conditions. This selection was important, since for measuring how similar clusters in different phases are, we employed the Adjusted Rand Index (ARI) metric (see Section 3) and including features which remain consistent regardless of the experimental phase could inflate the ARI score, potentially leading to false conclusions. This approach allowed us to derive ARI scores that genuinely reflected the similarity or diversity of participant responses to the tasks, providing insights into the consistency of physiological reactions across the participants.

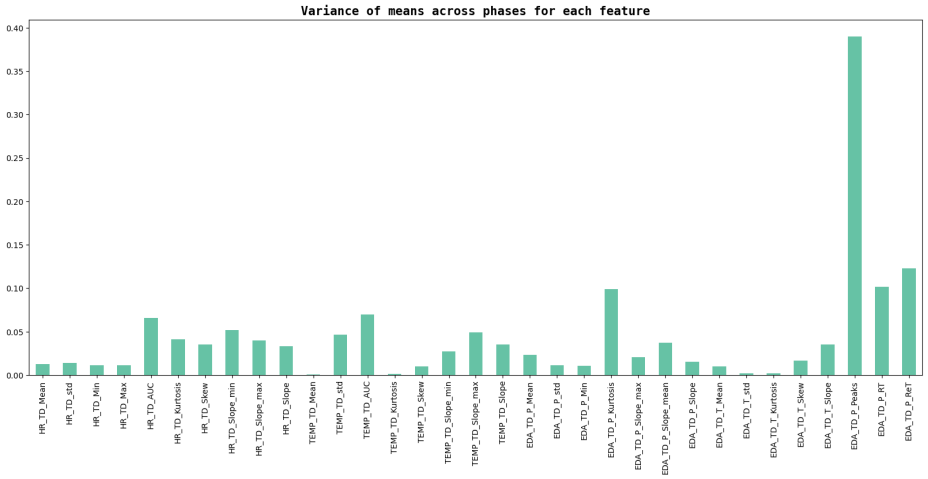


Figure 2: Variances of means for each feature across the different phases. Higher value suggests that the feature is reacting a lot to the different phases of the experiment

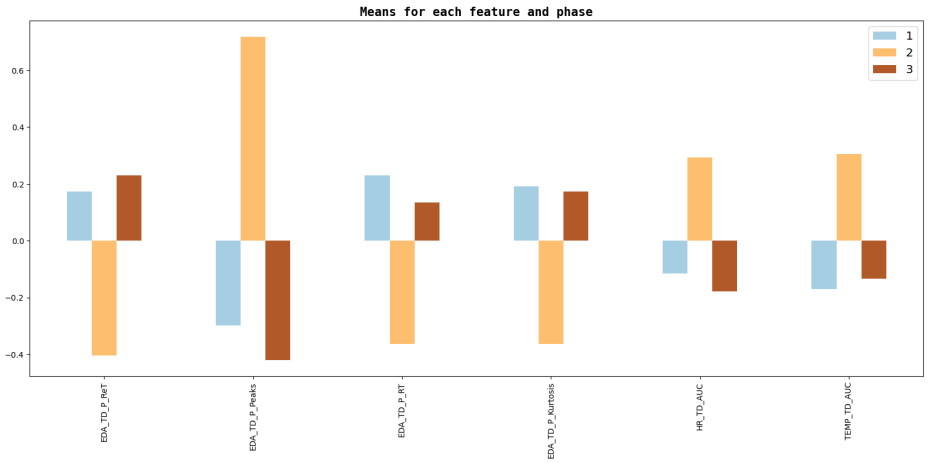


Figure 3: Means for 6 selected features for each phase respectively

This feature selection resulted in 6 attributes shown in Figure 3. The figure further underlines how the puzzle phase (yellow) substantially differs from the remaining two phases, suggesting that these features could also be good candidates for e.g. supervising tasks.

3 Methodology

Our objective was to determine if there was a clustering consistency among participants across different phases based on their physiological responses. To accomplish this, we decided to analyze the phases independently by clustering per round and per phase. This split was done because every participant had data from four rounds, and although some summary metrics such as the mean of those rounds could be calculated, we decided to keep the data separated to avoid losing information, making less assumptions and getting a more fine-grained view of the clusters, to for example see, if the evening rounds differ from the morning ones. The downside was that 12 separate clustering analyses had to be conducted. This also meant that we had to standardize each phase per round separately to avoid leaking information from one clustering to the other. However, this approach presented a challenge as this allowed only 26 data points available for clustering, and as mentioned in Section 2.2.2 clustering them in a large high dimensional space could be troublesome. We assessed the relevance of features for clustering based on their correlation, significance in terms of changes across phases, and silhouette scores. The silhouette score is a metric used to calculate the significance of a clustering technique. It ranges from -1 to 1, with 1 indicating well-separated and distinct clusters, 0 signifying indifferent clusters, and -1 suggesting that the clusters are assigned in the wrong way [1]. This approach ensured that only the most informative features were considered for clustering. Next, we applied PCA to reduce the dimensions even further, and then we performed clustering in this reduced space.

We chose to use K-means clustering due to its simplicity and widespread usage in similar analyses. Furthermore, given the time constraints of this assignment, K-means clustering offered a straightforward approach that could be implemented efficiently to achieve our research objectives. As described in Section 3.1, we needed to decide on the number of clusters for each round and phase. This was done using the elbow method. In this method we iterated from $K = 2$ to $K = 26$, and for each value of K we calculated the within-cluster sum of squares (WCSS), defined as the sum of the square distances between the centroids and each point. Then to identify the best number of clusters, we plotted K versus the corresponding WCSS values, and we selected the value of K at the point where the plot started to resemble an "elbow", which indicates the optimal number of clusters. [3]

3.1 K-Means Clustering

K-means clustering is an iterative descent method, that requires a pre-determined number of clusters. It uses the squared Euclidean distance as the dissimilarity measure, and aims to minimize the WCSS distances. It achieves this by assigning each of the N observations to one of the K clusters in a way that minimizes the average dissimilarity of the observations from the cluster

mean

$$W(C) = \sum_{k=1}^{K} N_k \sum_{C(i)=k} ||x_i - \bar{x}_k||^2, \tag{1}$$

where $\bar{x}_k$ is the mean vector associated with the kth cluster, and $N_k = \sum_{i=1}^{N} \mathbb{I}(C(i) = k)$. [5, p. 509] Note that there is no inherent ordering of clusters in (1), and that it seeks a local rather than a global optimum, meaning that the results can vary depending on the random initialization. Due to this sensitivity, it is important to execute the algorithm multiple times from various random initial states to ensure robustness and reliability in the clustering outcomes. [6, ch. 12.4.1]

3.2 Cluster Similarities - Adjusted Rand Index

To compare cluster similarities we will use the *adjusted Rand index* (ARI) – this is the Rand index adjusted for chance. [8] A pair is defined as two observations that have each been labelled under two different clusterings. The (regular) Rand index is computed as the ratio of pairs that are given the same label in both clusterings to the number of all pairs in the dataset:

$$\mathrm{RI} \triangleq \frac{C_a}{C_t}, \tag{2}$$

where C_a is the number of agreeing pairs and C_t is the number of possible pairs in the dataset. The Rand index is therefore a number between 1 (perfectly matching labels) and 0 (no pairs with agreeing labels).

As the Rand index can clearly be non-zero even for random labels, a modified metric called the adjusted Rand index (ARI) has been proposed. By adjusting with the expected Rand index of a random labelling,

$$\mathrm{ARI} \triangleq \frac{\mathrm{RI} - E\left[\mathrm{RI}\right]}{\max\left(\mathrm{RI}\right) - E\left[\mathrm{RI}\right]}, \tag{3}$$

the metric is close to 0 when random clustering has occurred – although it can now also become negative. [7]

One huge advantage with using the ARI metric is that it requires no assumptions on the cluster structure. An often cited disadvantage of the Rand index is that it requires ground truth labels. In our case, however, we use the metric to compare clustering across the phases of the experiment (instead of comparing to some absolute truth).

We will employ the ARI under the hypothesis that individuals that cluster together in one of the phases of the experiment also cluster together in the other two phases. Then the ARI will be a metric of to what extent this hypothesis is true.

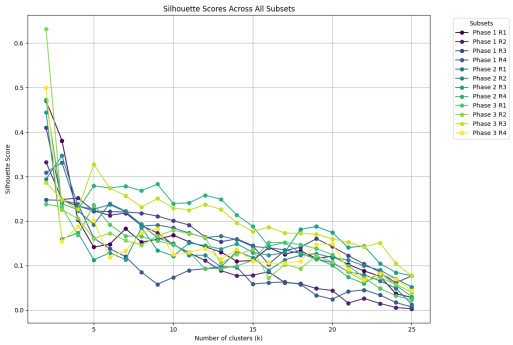
4 Results

As discussed in Section 3, we used the similarity score to asses the clustering quality, and the results of this can be seen in Figure 8 and 4. Looking at Figure 4, we observe that using PCA to reduce the number of features from six to two resulted in significantly improved silhouette scores. This was our main motivation behind using the dimensionality reduction method, as six dimensions still showed to be a bit too many for only 26 recorded observations. It's also important to mention that the two chosen principle components still capture 70% of the variance in the data.

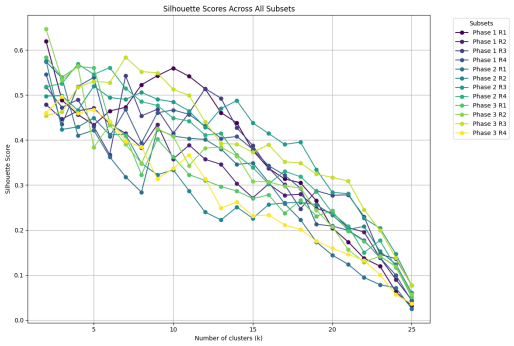
When $K = 6$ the average silhouette scores lie between 0.37 and 0.58, indicating that the clusters are reasonably well-separated and exhibit moderate to good compactness and cohesion. While not being exceptionally high, these silhouette scores indicate acceptable clustering quality, with clusters that are reasonably well-defined.

Then using these two features for clustering, we needed to estimate the optimal number of clusters K since we lacked a clear objective for the subgroups. For this task we used the elbow method and the results can be seen in Figure 5. We see an "elbow" at around $K = 6$, and combining this result with the average similarity score we can conclude that $K = 6$ is a suitable choice for the number of clusters, as it represents a balance between cluster separation and cohesion, thereby providing meaningful grouping of the data points. For a more detailed description of the number of clusters chosen, we refer to Table 2.

Given our small dataset of only 26 participants, selecting the right features was paramount, as the ARI metric proved to be highly sensitive to both the number of clusters chosen for each phase



(a) Clusters created using K-means clustering with features that are strongly dependent on the experimental phase.



(b) Clusters created using K-means clustering on the two first principal components of features that are strongly dependent on the experimental phase.

Figure 4: Average silhouette scores as a function of the number of clusters $K \in \{2, 3, \ldots, 26\}$.

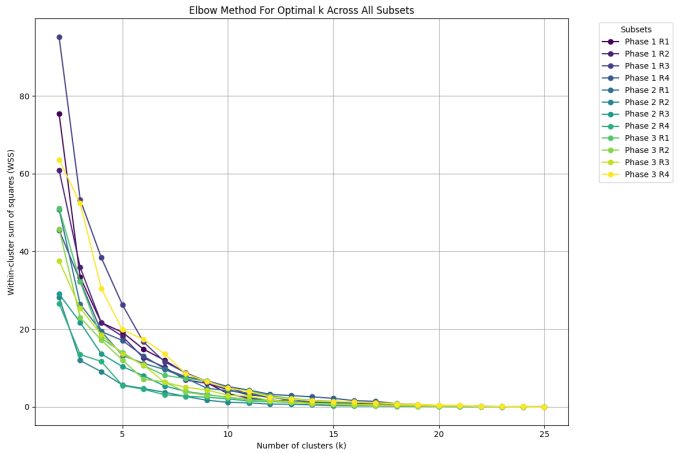


Figure 5: Clusters created using K-means clustering with features that are strongly dependent on the experimental phase where PCA was applied to reduce the dimensions to two.

per round and the number of features considered. The features strongly dependent on the experimental phases, provide insights into how the participants respond to stressful situations, such as solving puzzles. However, clustering based solely on these dynamic features revealed that despite substantial fluctuations in physiological responses, no discernible pattern emerged indicating participants' cohesive transitions from one phase to the next as seen in Figure 6. Interestingly, when clustering on features less dependent on the phases, we observed a higher ARI score. This underscores the importance of including only those features directly affected by experimental conditions, as incorporating unaffected features could artificially inflate the ARI score, potentially leading to false conclusions. Therefore, by clustering separately and prioritizing dynamic features, we ensure a more accurate assessment of participants' responses to the experimental challenges.

4.1 Emotional correlation

In the dataset we were also given data from an emotion-based survey that participants had to fill before the beginning of each phase. An obvious second question that came to our minds was whether or not there was a correlation between the answers people gave in these surveys and which clusters they were put in. As seen from the results above, the clusters did not hold any information about how people would react in the future/have reacted in the past, but maybe they held some information about what emotion the participants in them were feeling.

To measure this, we calculated the global mean score the participants had given the emotions for each round/phase combination and compared that to the mean score in each cluster. We then added up the absolute value of these differences across all clusters and all rounds to find out which emotion stood out the most in clusters. We quickly ran into the issue that some clusters only had

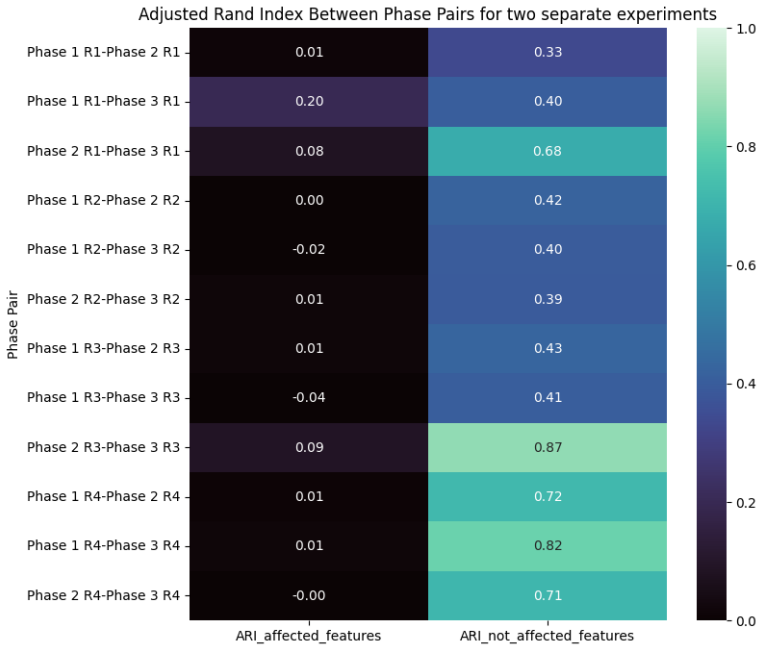


Figure 6: Clustering on features highly affected by experiment conditions (see section 2.2.2) results in participants not clustering similarly across phases as indicated by the low ARI scores (left). For contrast, clustering on features that are not significantly affected by the experiment conditions shows that participants tend to move together across the different experimental phases and co-appear together (right).

a single participant, meaning the deviation would show itself to be really large. This doesn't really answer whether or not there is a correlation between clusterings and emotional responses as much as a large cluster with slightly smaller deviation would. To compensate for this, we introduced a weighted deviation method where we multiplied each cluster's deviation from the mean by the number of participants in said cluster. The mathematical formulation of this concept is given by the following equation:

$$\text{Weighted Deviation}(\text{Phase}, \text{Emotion}) = \sum_{\text{Rounds}} \sum_{\text{Clusters}} n \times |E_1 - E_2|$$

where:

n: The number of participants in each cluster.

E_1: The global mean of the emotion across all data points in the phase, providing a baseline for comparison.

E_2: The mean of the emotion within a specific cluster.

This formula helps normalize the influence of each cluster based on its size, ensuring that smaller clusters with major deviations do not disproportionately influence the results, thus providing a more accurate picture of how emotions vary across different clusters within each phase.

We found that there was a clear difference between the emotions. Positive emotions such as "Active", "Inspired" and "Determined" all had a high deviation score while negative emotions such as "Ashamed", "Hostile" and "Afraid" all ranked quite low. Although promising, it is also important to note that if these scores had a very low variance to begin with, them ranking low in weighted deviation was expected. This low variance for negative emotions was possibly caused by the fact that the phases didn't introduce any factor which would lead to the participants being for example afraid. To further analyse this we also calculated the weighted deviation scores for each emotion for random clusters and compared the results.

Figure 7 reveals, that the graphs look quite similar for the K-means and random clusters. One could argue that for the final phase, the K-means clusters are somewhat related to the positive emotions, but the difference is not large enough for this conclusion to be made.

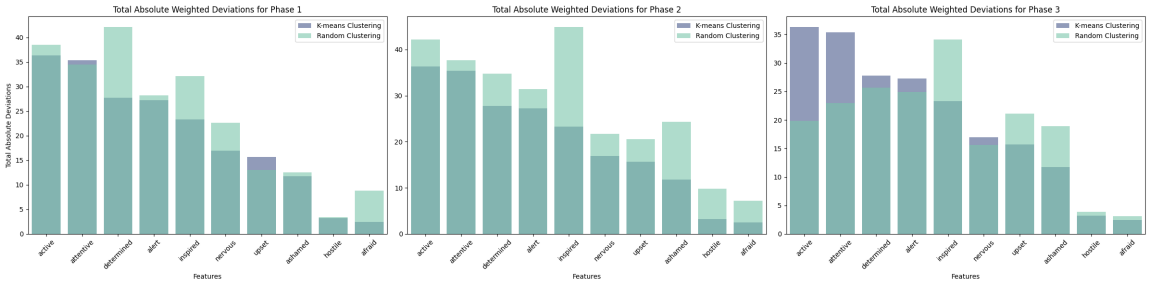


Figure 7: Graph showing the Weighted Deviation scores of the different emotions for both K-means and random clustering.

5 Conclusion

In this project we did an in-depth analysis of physiological responses across different phases of an experimental setup [2], utilizing the EmoPairCompete dataset to investigate whether participants' physiological signals cluster consistently across these phases. Through careful data preprocessing, feature selection, and the application of K-means clustering and Adjusted Rand Index, we aimed to identify patterns that would suggest common physiological reactions under controlled experimental conditions.

Our analysis highlighted the importance of selecting features that exhibit significant variability across different experimental phases. This approach enabled us to focus on changes that are most likely to reflect true physiological responses to the experimental tasks rather than background noise or invariant aspects of the data. The findings underscored the sensitivity of the Adjusted Rand Index (ARI) to the choice of features and the number of clusters. By focusing on features that varied significantly with experimental conditions, we ensured that our analysis provided a more genuine reflection, avoiding the inflation of similarity scores that could have resulted from less responsive features.

The clustering analysis revealed that while some features indicated a strong response to experimental conditions, the overall clustering did not show consistent groupings of participants across phases. This suggests that individual physiological responses to the experimental conditions may be more varied than hypothesized, indicating a complex interplay of emotional and physical factors that does not necessarily align into stable cluster patterns across phases.

In the study we were presented with the complexity of physiological data and the challenges involved in clustering such data in a meaningful way. Further, the small number of participants limited the generalizability of the findings. Future studies could benefit from a larger dataset to validate the observed patterns and refine the clustering methodology. While we managed to reduce the feature space, the impact of this reduction on the clustering outcome suggests that further refinement in feature selection and dimensionality techniques could enhance the analysis. Future work might explore alternative clustering algorithms or other machine learning models that could handle the high dimensionality and variability of physiological data more effectively.

In conclusion, by continuing to explore these complex datasets with refined tools and techniques, researchers can better understand the nuanced ways in which individuals react to different stimuli, potentially leading to more personalized approaches in health monitoring and treatment strategies.

References

[1] Ashutosh Bhardwaj. *Silhouette Coefficient: Validating Clustering Techniques*. 2020. URL: https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c.

[2] Sneha Das et al. "EmoPairCompete - Physiological Signals Dataset for Emotion and Frustration Assessment under Team and Competitive Behaviors". In: *ICLR 2024 Workshop on Learning from Time Series For Health*. 2024. URL: https://openreview.net/forum?id=BvgAzJX40Z.

[3] *Elbow Method for Optimal Value of K in K-means*. 2023. URL: https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/.

[4] Giorgos Giannakakis et al. “Review on Psychological Stress Detection Using Biosignals”. In: *IEEE Transactions on Affective Computing* 13.1 (2022), pp. 440–460. DOI: 10.1109/TAFFC.2019.2927337.

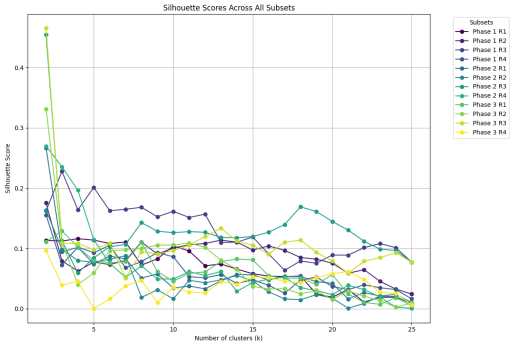
[5] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001.

[6] Gareth James et al. *An Introduction to Statistical Learning with Applications in Python*. Springer Texts in Statistics. Cham: Springer, 2023. ISBN: 978-3-031-38746-3. DOI: 10.1007/978-3-031-38747-0. URL: https://link.springer.com/book/10.1007/978-3-031-38747-0.

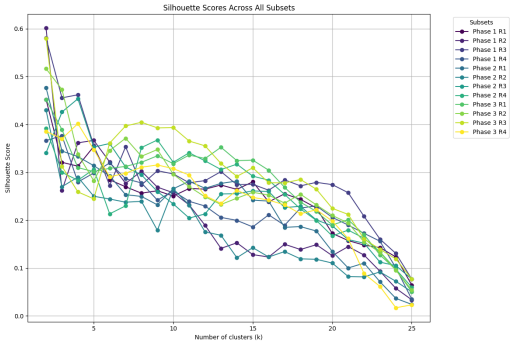
[7] Scikit-learn contributors. *Rand index, Scikit-learn user guide — Scikit-learn Python machine learning library*. https://scikit-learn.org/stable/modules/clustering.html#rand-index. [Online; accessed 29-April-2024]. 2024.

[8] Wikipedia contributors. *Rand index — Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/w/index.php?title=Rand_index&oldid=1171913170. [Online; accessed 29-April-2024]. 2023.

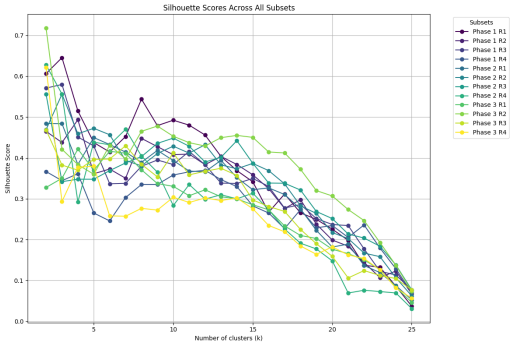
A Additional Results



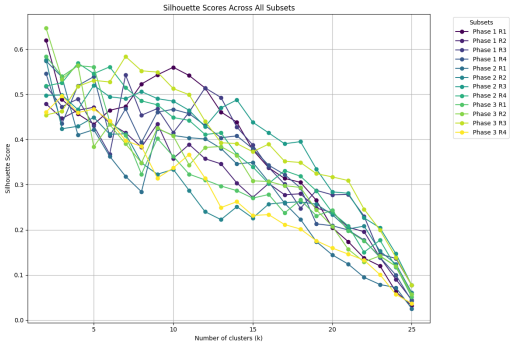
(a) Silhouette scores for clusters with all features having correlation scores below 0.9, while retaining only one of the highly correlated features.



(b) Silhouette scores for clusters created using K-means clustering with just the means of the biosignals.

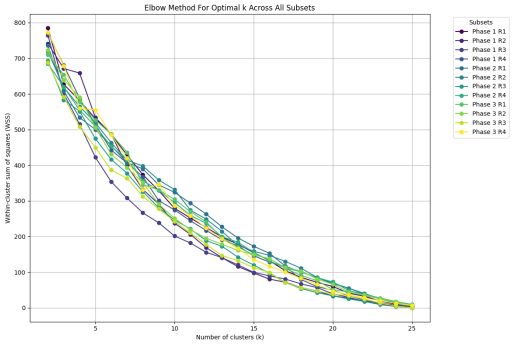


(c) Silhouette scores for clusters created using K-means clustering with features that are strongly dependent on the experimental phase.

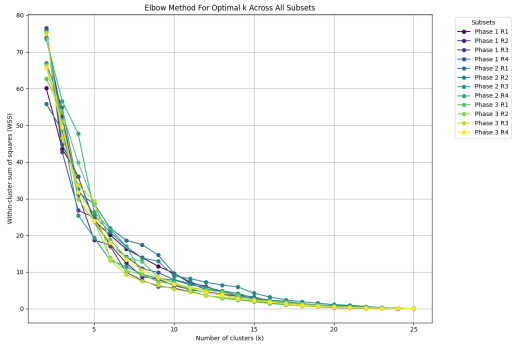


(d) Silhouette scores for clusters created using K-means clustering with features that are not reacting to the experiment conditions

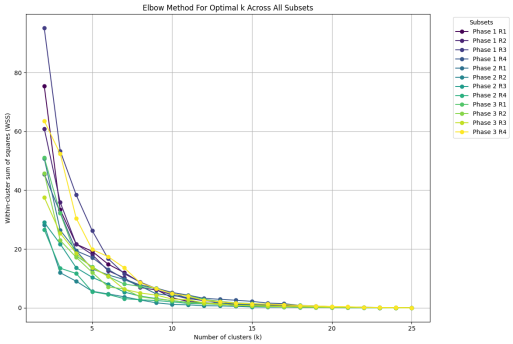
Figure 8: Average silhouette scores as a function of the number of clusters $K \in \{2, 3, \ldots, 26\}$.



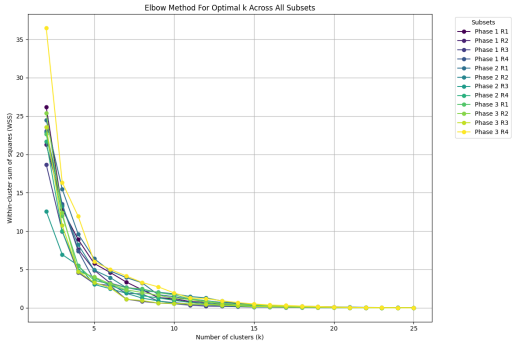
(a) Clustering using all features with correlation scores below 0.9, while selectively retaining only one of the highly correlated features.



(b) Clustering with just the means of the biosignals.



(c) Clustering on the two first principal components of features that are strongly dependent on the experimental phase.



(d) Clustering on features that are not reacting to the experiment conditions

Figure 9: Within-cluster dissimilarity as a function of the number of clusters $K \in \{2, 3, \ldots, 26\}$.

Phase	Round	Optimal Clusters
1	1	5
1	2	5
1	3	7
1	4	6
2	1	4
2	2	4
2	3	4
2	4	8
3	1	7
3	2	6
3	3	6
3	4	5

Table 2: Optimal Number of Clusters for Each Phase and Round.