02582 Computational Data Analysis

Case 2

Carlos de Santiago (s222766)
Laura Figueiredo Tor (s222797)
Zhijian Feng (s213122)
Dimitrios Loukas (s220514)

1 Introduction

The aim of our project was to conduct an exploratory (unsupervised) analysis of the biosignal data collected from the Empatica E4 wristband in a controlled experiment based on solving a puzzle in teams of two, in order to discover interesting patterns, trends, or relationships within the data that can lead to further research.

Data description

This dataset includes physiological measurements such as electrodermal activity (EDA), heart rate (HR), temperature (TEMP), and blood-volume pulse (BVP) collected from 26 individuals over four acquisition rounds. The biosignals were collected over the 3 phases that consist of a unique round of the experiment: a resting period, an emotion-eliciting condition (puzzling phase), and a recovery period. The dataset also includes self-rated questionnaires (I-PANAS-SF) completed by participants after each phase. This questionnaire includes five positive emotions (active, alert, attentive, determined, inspired) and five negative emotions (afraid, ashamed, hostile, nervous, upset), that are rated on a scale from 1 to 5. Frustration and difficulty levels were also assessed using a scale from 0 to 10, while the latter was conducted only after the puzzle phase.

The original dataset consists of separate .csv files for each participant for each phase of each four rounds. The data for each phase is stored in 5 files, 4 of them contain one of the individual biosignal measurements (TEMP, HR, EDA, and BVP) and the sampling time, while the last one contains the responses to the questionnaires for this phase (upset, hostile, alert, ashamed, inspired, nervous, determined, attentive, afraid, active, frustrated, difficulty) and additional information about the participant and team ID, the model of the wristband and whether the person is a puzzler (solver) or not (instructor). Finally, all the aforementioned files were merged and loaded into a single data frame for further pre-processing and data manipulation.

2 Pre-processing and data manipulation

This section provides an overview of the preprocessing steps performed to clean and transform the raw data into an appropriate format for further analysis. Specifically, we will describe the procedures used to remove missing data, find outliers and normalize the data.

At first, a new *participant* feature was created by merging the directory names *D*, *Sub_D*, and *ID*, in order for each participant to have a unique identifier (as the IDs provided in the response files are not unique for each individual). The *Sub_D* feature was created due to the data recompilation in cohort 3. Furthermore, we discovered that even though the response to the *difficulty* question was only needed after the puzzle phase, there were participants with responses to that question in phases other than *Phase 2*, so these values were considered to be outliers and were set to 0. However, there were other participants without any information about the *difficulty* in the other phases, so we decided to add a value of 0 in that cases to ensure that all participants have 0 values on the *difficulty* in phases 1 and 3 from the data frame.

To ensure consistency in the biosignals data, *TEMP*, *HR*, *BVP* and *EDA*, it was important to have measurements for each participant taken over the same timeframe, which was supposed to be 5 minutes for each phase. Although we aimed to have a consistent recording time of 5 minutes for

each phase, the recording times varied as participants were responsible for starting and stopping the recording for each phase. Therefore, in some cases, the recording time was less than 5 minutes, making it impossible to obtain the ideal measurement timeframe. To address this issue, we decided to set a minimum accepted timeframe of 4.5 minutes (270 seconds), and any case with a lower timeframe was discarded from the analysis. Eighteen individual cases inside the data frame with all the rounds, phases, and participants included (312 rows) were deleted due to this.

Additionally, each physiological signal was recorded with a different sampling rate, resulting in varying total numbers of measurements. To address this issue, we calculated the sampling rate per second and resampled the data accordingly, ensuring that there was one measurement per second for each signal. One of the problems when working with these time-series variables is that we cannot ensure that all participants started and finished recording their physiological variables at the same time. For this project, we have ignored this problem and considered the same timeframe of 270 seconds for each phase.

Next, we performed normalization of the physiological data (applying the *normalize* function implemented in *scikit-learn* python library) by dividing each value by its Euclidean norm. In order to normalize the physiological data, we employed a column-wise normalization individually for each physiological variable. The feature columns, for a physiological variable, consisted of a list of length 270 for each observation (294 rows taking into account the 18 observations previously removed). Therefore, we normalized the data column-wise for each second of the recordings. This means that we took all the measurements of a specific second from all the observations in a given feature column, and then we divided each measurement by the Euclidean norm of all the measurements of that specific second across all observations. This allowed us to scale each feature consistently, making them comparable across all observations.

Later, other processing steps were carried out for the different methodologies applied. For example, the creation of a time-window data frame with the mean and standard deviation from the physiological variables every 15 seconds for each participant was completed for analyzing the behavior of the participants from the experiment. This type of data frame was used for some of the principal component analysis methodologies implemented in this project, which are explained in more detail in section 3.

Moreover, some of the participants in selected phases have missing values in some of the questions from the *response* data. For fixing that issue, we replace the missing values according to the mean answer for that question in the other rounds for that phase in that specific individual. Finally, one hot encoding was performed in the feature *team_ID*, as the numerical value of the variable does not carry any meaningful information, because there is no inherent numerical order or relationship between the different team IDs.

3 Unsupervised methodologies

3.1 Clustering methods

3.1.1 K-means clustering for variables inspection

We deemed it possible that there was some relation between the physiological variables (*TEMP*, *HR*, *BVP*, and *EDA*) and the questionnaire responses. For instance, we hypothesized that a high value for emotions such as "afraid", "nervous", or "upset", could correlate with a high value for the heart rate.

To asses this hypothesis, we performed K-means clustering and Spectral clustering (SC) for all the participants in two situations: first taking into account only the physiological variables, and later including exclusively the questionnaire responses. We chose these two methods because they are both suitable for unsupervised learning. This procedure aimed to analyze the clusters formed by the participants in both situations and determine whether the same clusters were formed based on the physiological variables and the questionnaire responses. In other words, our objective was to analyze the clusters formed in both situations to test whether the physiological variables and the response questionnaire give the same type of information.

For the first part of the analysis, we built a matrix in which the rows were the observations, and the columns the mean and the median for the normalized values of the physiological variables taking a time-step of 15 seconds (as noted in section 2). That is, for each of the 4 physiological variables the 270 normalized measures were grouped into sets of 15, and the mean and median of these sets were used as variables to conduct the clustering. We also considered including the standard deviation of each group, but it considerably increased the dimensionality of the data and did not yield better results for the clustering, so we considered it as not informative and discarded this option. Thus, the clustering was conducted on a matrix with 294 (observations for the different participants across the different rounds and phases) and 144 columns ($\frac{270}{15} \cdot 4 \cdot 2$). For the second part of the analysis, we used a matrix with 294 rows and 12 columns, one for each emotion addressed in the questionnaire. 7 NaN values were found and replaced with the mean of the column to have the least possible impact on the results.

3.1.2 Experiment phases analysis

Similar to the procedure that is explained in the following section (3.2.2) for the participant's behavior analysis, we wanted to analyze possible clustering according to the phases of the experiment. In this way, we could try to conclude if there is some kind of characteristic behavior depending on the phase of the experiment. In this manner, we treated the data the same way as explained in the section 3.2.2 but without joining all the corresponding phases for each participant (only joining all the rounds for each participant). Therefore, we ended up with a time-window data frame of 26·3 rows.

3.2 Principal Component Analysis and Self-Organizing Maps

3.2.1 Time series analysis

In this part, we aimed to analyze biosignal data using principal component analysis (PCA) as our primary method for time series analysis. More specifically, we investigated the behavior of physiological data, *TEMP*, *HR*, *BVP* and *EDA*, by performing PCA on the biosignal data of each participant and for each round separately. Our goal was to concatenate all three phases and identify potential patterns in the behavior of principal components across the three phases, comparing all participants. To achieve this we transformed the normalized data into a new data frame for each participant, where the rows were the measurements sampled per second during all 3 phases (total of 810 seconds) and the columns were *Time*, *TEMP*, *HR*, *BVP* and *EDA*. We then performed PCA for each participant, excluding the *Time* feature, and analyzed the results.

3.2.2 Participants behaviour analysis

In this part, we wanted to find some type of clustering, patterns, or outliers using PCA for each individual's physiological and response measures. This way, we constructed a time-window data frame with the mean and standard deviation for each physiological variable of each participant every 15 seconds and joint it to the responses of each individual for each round and phase. Then, we joined all the rounds for each participant by doing the mean of those rows, followed by joining all the corresponding phases for each participant by doing the mean of those rows. By doing this procedure, we ended up with a time-window data frame of 26 rows (1 row per participant), so we could analyze possible patterns in the behavior of some individuals according to whether they were in a specific team or by the task they performed in the experiment (puzzler solver or instructor).

Additionally, we tried to implement this time-window data frame of 26 rows in self-organizing maps (SOM) to compare the PCA results with a spatial representation provided by implemented SOM methodologies. This was achieved by using *minisom* library in Python.

4 Results

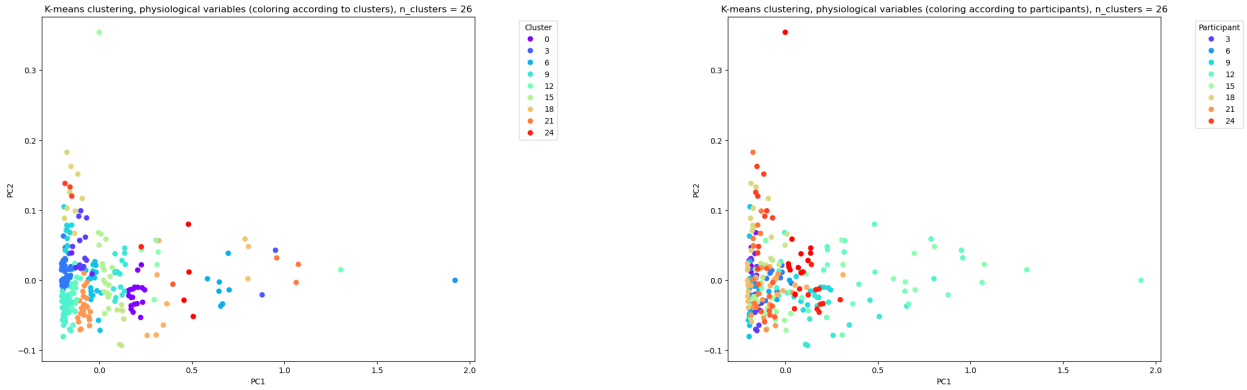
4.1 Clustering

4.1.1 K-means clustering for variables inspection

Before performing K-means clustering, we calculated the Gap statistic to determine the optimal number of clusters. The results were not promising, as they concluded that the optimal number of clusters for the matrix containing the physiological variables' information was 294, the number of observations, and 265 for the matrix containing the questionnaire variables, also considerably high. This may suggest that the data is not well-suited for clustering.

Nonetheless, we proceeded with the exploratory analysis. First, we set 26 as the number of clusters, the number of different participants, to observe if the 26 formed clusters resembled the 26 participants across the different phases. As shown in Figures 1 and 2, the clusters obtained were not clearly separated from one another (at least when represented in 2 dimensions), and the clusters did not correspond to the different participants. Thus, we could conclude that the participants did not have physiological variables or questionnaire responses that made them different from each other.

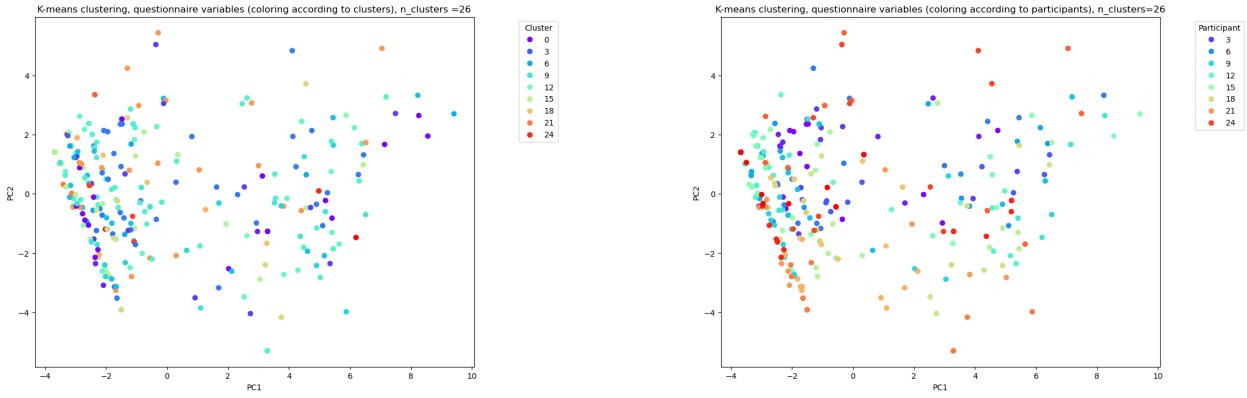
In addition, we compared the clusters formed taking the physiological variables into account and the clusters formed when including only the questionnaire variables by calculating the Adjusted Rand Index (ARI). ARI is a measure of the similarity between two clusterings of the same data. It takes a value between -1 and 1, where a value of 1 indicates perfect agreement between the two clusterings, a value of 0 indicates random clustering and a value of -1 indicates complete disagreement between the two clusterings. In this case, the value obtained was 0.0232, which confirms that the observations cluster differently when taking into account the two sets of variables.



(a) Observations colored by the K-mean clusters

(b) Observations colored by participants

Figure 1: (a) K-means clustering with $n_{clusters} = 26$ based on **physiological variables** vs (b) different participants. The observations are plotted on the first two principal components.



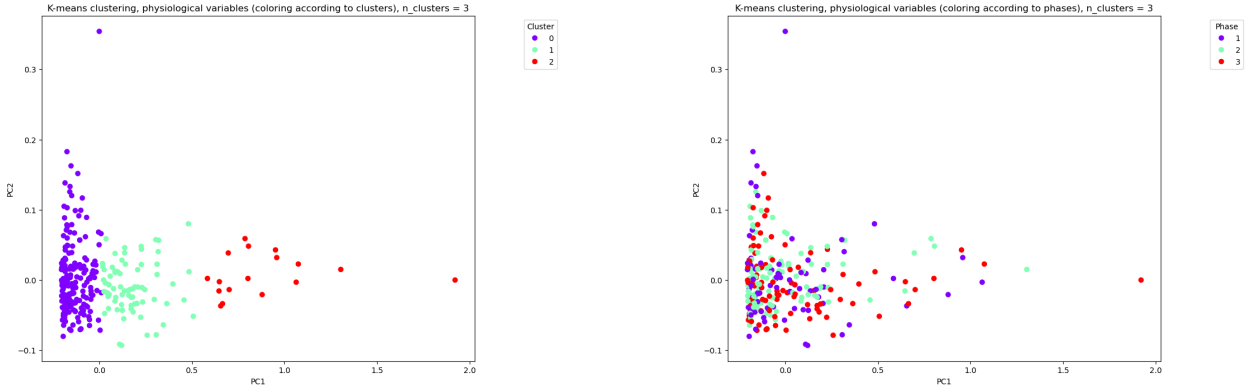
(a) Observations colored by K-mean clusters

(b) Observations colored by participants

Figure 2: (a) K-means clustering with $n_{clusters} = 26$ based on **questionnaire variables** vs (b) different participants. The observations are plotted on the first two principal components.

After this first clustering analysis, we performed the same analysis but setting the number of clusters

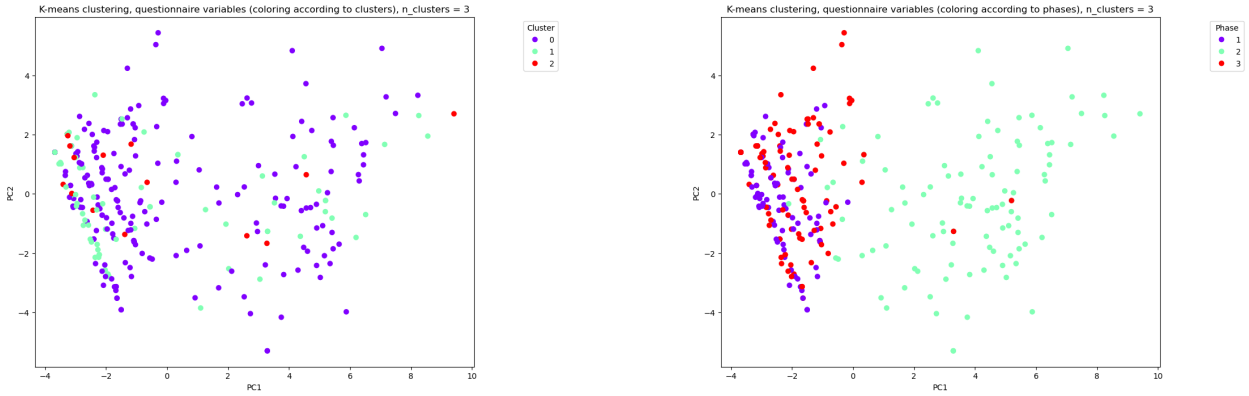
to 3 (number of different phases), to determine if the physiological or questionnaire variables made the observations cluster depending on the phase they were taken in (Figures 3 and 4). As can be observed, the observations were not clustered by the K-means algorithm according to the phase, either.



(a) Observations colored by K-mean clusters

(b) Observations colored by phases

Figure 3: (a) K-means clustering with $n_{clusters} = 3$ based on **physiological variables** vs different phases. The observations are plotted on the first two principal components.



(a) Observations colored by K-means clusters

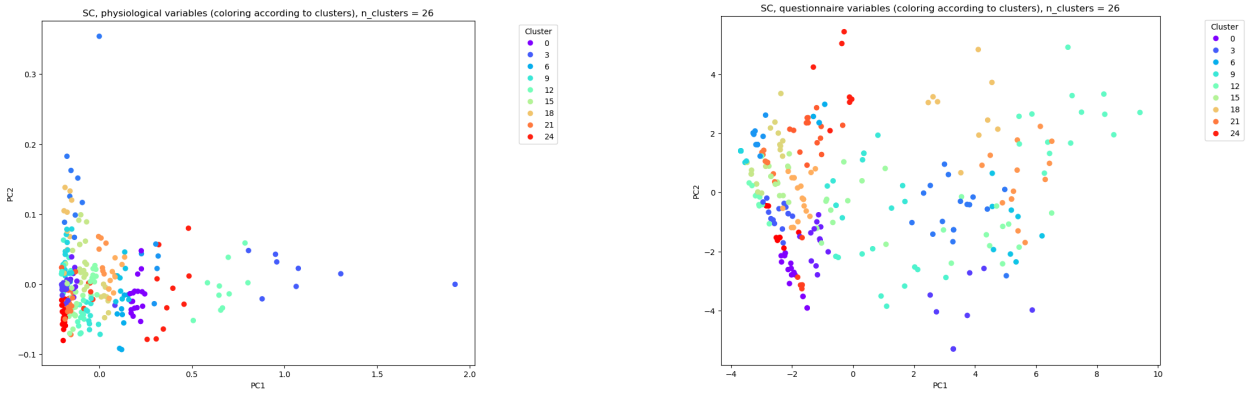
(b) Observations colored by phases

Figure 4: (a) K-means clustering with $n_{clusters} = 3$ based on **questionnaire variables** vs (b) different phases. The observations are plotted on the first two principal components.

Moreover, in this case the calculated ARI had a negative value, -0.011, indicating that the clusters obtained when taking into account to the physiological differed in great measure to the clusters obtained when taking into account the physiological variables.

In order to explore if other methods of clustering would yield better results, we also performed SC. As SC takes into account the similarity of data points in the high-dimensional space, in some cases it can perform better than K-means, specially when the data has a non linear structure or when the clusters are not clearly separated in the high-dimensional space.

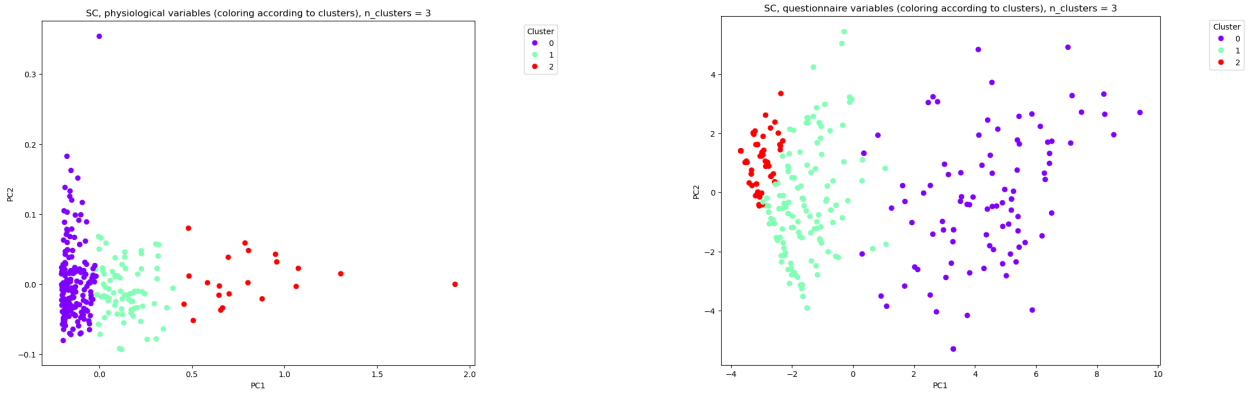
Figures ???5 and ???6 show the results of performing SC (calculating the similarity with K-nearest neighbors method, with 10 neighbors) for 26 and 3 total number of clusters, as did in section ???4.1.1. As can be observed, the results were apparently not better than those obtained with the K-means algorithm. Indeed, the calculated ARIs were 0.02611 for 26 clusters and 0.03746 for 3 clusters. Even if both were positive this time, the values are very close to zero, indicating no similarity between the clustering according to physiological variables and the one according to questionnaire variables.



(a) SC based on PV, colored by clusters

(b) SC based on QV, colored by clusters

Figure 5: SC with nclusters=26n$_{clusters}$ $= 26, using K - nearest neighbors to calculate simmilarity (K = 10). PV refers to "physiological variables", QV to "questionnaire variables". The observations are plotted on the first two principal compon$



(a) SC based on PV, colored by clusters

(b) SC based on QV, colored by clusters

Figure 6: SC clustering with $n_{clusters} = 3$, using K-nearest neighbors to calculate simmilarity (K=10). PV refers to "physiological variables", QV to "questionnaire variables". The observations are plotted on the first two principal components.

Several values for different parameters were tried to ascertain that SC could not yield results that led to different conclusions. The number of neighbors was changed, as well as the similarity measure (the Radial Basis Function with different values of γ was also tried), but the overall conclusion did not change: the observations do not cluster according to the individual participants or the phases, and the physiological variables and questionnaire variables lead to the formation of different clusters.

4.1.2 Experiment phases analysis

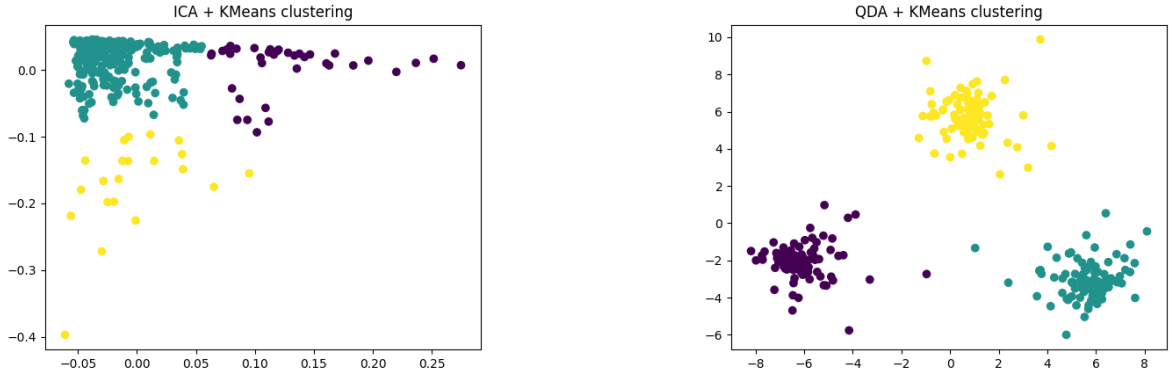
In this section, we expected to separate observations belonging to different phases as much as possible through technology. The reason we did this was that different participants underwent the same experience, and this experience was divided into three different stages, which we can intuitively get to based on life experience when you are before, during, and after playing poker or facing a puzzle. The physiological indicators of participants were different, and we understood that the basic situation was the same for each participant. Based on this assumption, we wanted to aggregate the samples at different stages as much as possible through different methods.

In data preprocessing, we still used the method described in sections 2 to collect data over a period of time at certain time intervals. In this experiment, we performed feature statistics for BP, HR, EDA, and TEMP attributes, calculated the mean and variance of each time interval with a window size of 5, and generated the corresponding features. A final dataset size of (294, 433) was generated.

In order to better display the data at different stages, we initially performed ICA on the dataset, aiming to treat all attributes as multiple mixed signals and then decompose them into multiple independent signals, hoping that the data at different stages could thus be well distinguished. The experimental results with K-Means are shown in Figure 7a.

Through the experimental results, we could state that ICA can distinguish the samples of different stages to some extent, and we could know that the samples of different stages have different distributions. Based on these results, we used QDA to continue the optimization.

The reason for using QDA is that according to the results of ICA, we assumed that the samples at different stages would obey multivariate positive-terminus distribution, we hoped to map the samples to the new feature space, and we hoped that the samples of the same category would be as close as possible after projection, while the samples between different categories would be separated as much as possible after projection, so as to get better clustering effect. The experimental results with K-Means are shown in Figure 7b. After using QDA for mapping, the results of K-Means are significantly improved.



(a) Cluster result of ICA and K-Means

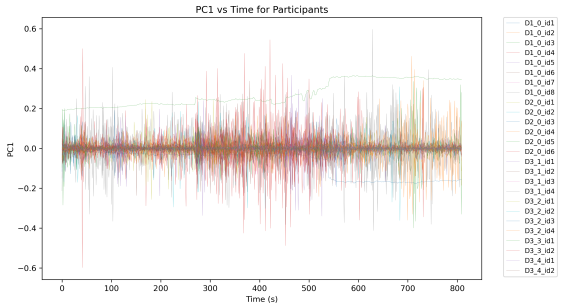
(b) Cluster result of QDA and K-Means

Figure 7: Cluster results for all participants with all the individual rounds joined as time-window data, in order to analyse possible clustering according to the 3 phases of the experiment.

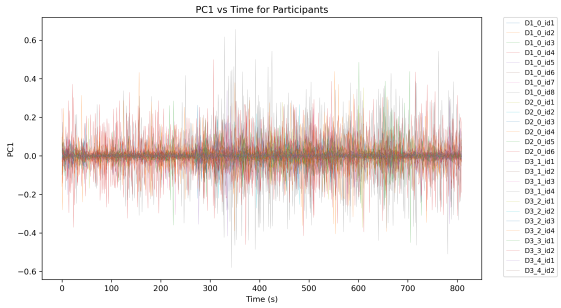
4.2 PCA

4.2.1 Time series analysis

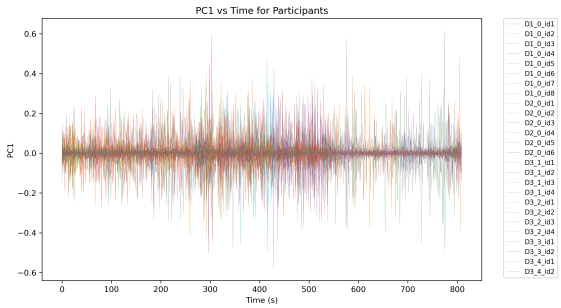
In the pre and post-puzzle phases, we expected the PC1 scores to remain stable because the participants were just resting and sitting calmly for 5 minutes. We expected them not to experience any significant levels of stress or frustration during these phases, resulting in the absence of significant changes in their physiological responses during the first and last 270 seconds. In contrast, in the Puzzle phase, we expected to see a significant increase in PC1 scores due to the stress and frustration due to the competition, perhaps followed by a decline as participants became more and more fatigued. However, our analysis showed that PC1 scores fluctuated around 0 throughout all three phases (Figure 8), indicating that there was no significant improvement or decline in performance over time. This unexpected result suggests that factors other than stress and familiarity with the task and technology may have influenced participants' performance.



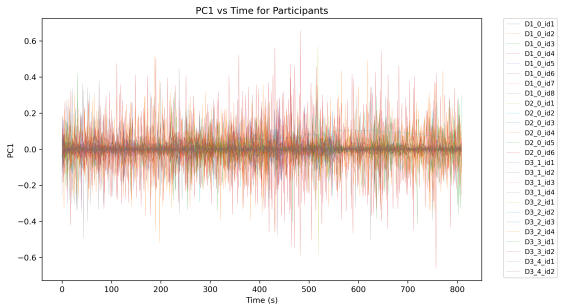
(a) PC1 vs Time for acquisition round 1 for all participants



(b) PC1 vs Time for acquisition round 2 for all participants



(c) PC1 vs Time for acquisition round 3 for all participants



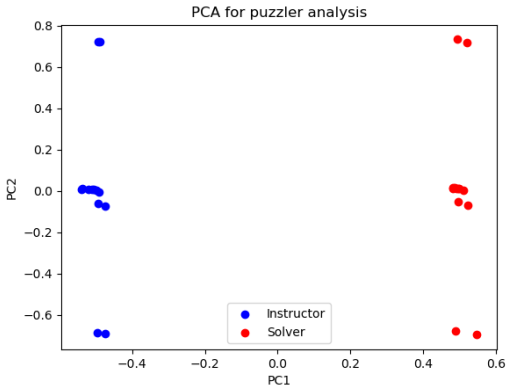
(d) PC1 vs Time for acquisition round 4 for all participants

Figure 8: Time series plots of PC1 for all participants for all four acquisition rounds. The x-axis represents time in seconds, while the y-axis shows the values of PC1, which is a linear combination of the original biosignal measurements. Each plot corresponds to a different acquisition round, and the lines represent the PC1 values for each participant.

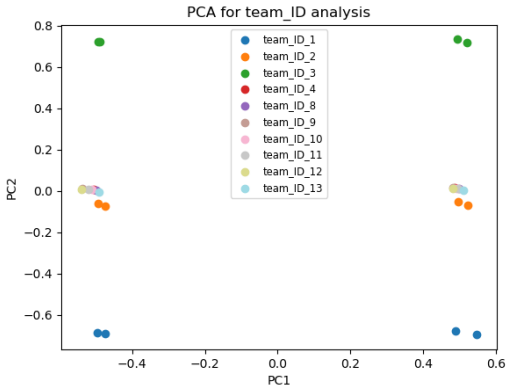
4.2.2 Participant behavior analysis

Joining all the rounds and phases data for each participant in a time-window data frame allowed us to search for patterns between some types of participants by implementing PCA. At first, we tried to look for some type of outlier behavior if the majority of participants clustered together in some kind of way, but we could not find this type of outlier behavior.

After that, we investigated whether the puzzle solvers clustered all together differently from the puzzle instructors, suggesting that the role achieved during the experiment is one of the main forces that drive the behavior of the participants. We could clearly observe a different clustering depending if an individual was a solver (puzzler = 1) or an instructor (puzzler = 0), as shown in 9a. In figure 9b, it is remarkable to notice how the majority of teams clustered all together in the same PC2 value, concluding that most of the individuals behaved in the same way despite the time the experiment was run, the difficulty of the puzzle, or with who was paired to complete the task. It also seems like the team distribution also has some kind of impact on the participant's behavior, as the individuals in the same team share practically the same PC2 values as observed in 9b.



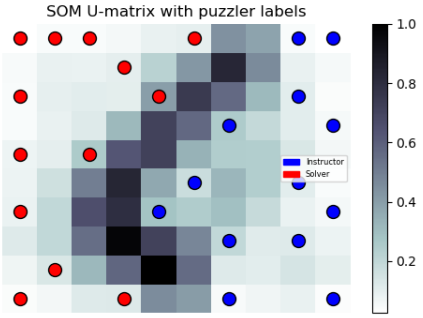
(a) PCA representation for the two first principal components, with solver individuals colored as red and instructor individuals colored as blue



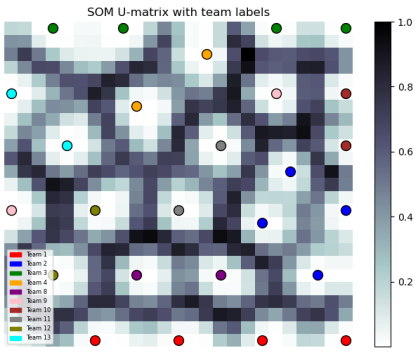
(b) PCA representation for the two first principal components, with each team represented with different colors

Figure 9: PCA plots for all participants with all the individual rounds and phases joined as time-window data. The x-axis represents PC1 and the y-axis represents PC2 values for each individual.

We used self-organizing maps (SOM) to help us better understand how participants clustered based on whether they were puzzle solvers or instructors and their team ID. The SOM clearly separated neurons related to puzzle solvers from those related to instructors (Figure 10a). However, when we included team information, the map was harder to interpret, as participants from the same team were positioned closer together but not perfectly correlated (Figure 10b). Nevertheless, we can conclude that people's feelings and attitudes were coordinated during cooperative tasks.



(a) SOM representation of 10x10 dimensions, with solver individuals colored as red and instructor individuals colored as blue



(b) SOM representation of 25x25 dimensions, with each team represented with different colors

Figure 10: SOM results from the U-matrix for all participants with all the individual rounds and phases joined as time-window data. The squares in the plot represent neurons, and the legend bar represents the distance between neighbor neurons, meaning that longer distance values indicate more dissimilarity between the neighbor neurons.

5 Conclusions

The time series analysis using PCA produced unexpected results, with PC1 scores fluctuating around 0 for all participants in all 4 acquisition rounds. Although some variation was observed, especially in rounds 1 and 3, where a slight increase in fluctuation was seen during phase 2, this was consistent with our hypothesis that higher stress levels during the puzzle phase would lead to higher PC1 scores. However, in acquisition rounds 2 and 4, the fluctuations were more evenly distributed across the phases, suggesting the existence of other factors that may be influencing participants' physiological responses beyond the stress of the puzzle phase. This hypothesis is reinforced by the fact that the physiological variables and the questionnaire variables led to the formation of very different clusters both when performing K-means clustering and SC. Overall, the unexpected results suggest that there may be additional hidden factors affecting participants' responses to the task beyond what we originally hypothesized.

By comparing all individual's behavior together, we can conclude that individual behavior is mainly determined by the role assigned in the cooperative task rather than the difficulty of the puzzle or other factors. Additionally, individual behavior can be also affected by the person you are coupled with when doing the experiment task. Both statements conclude that the way a person performs on a task is more influenced by the person who is coupled with and the role that is assigned for a task, rather than the task that person has to perform.

According to the experiment phases analysis of instructors and solvers, we can see that there are some differences in the physiological indicators of both, but in general, the physiological indicators of instructors and solvers are still very similar in each phase. Although they were responsible for different tasks, they both participated in the experiment at the same time, so their responses to the results were the same.