Danmarks Tekniske Universitet

DTU

Computational Data Analysis Assignment 2

Authors

Lydia Kasapi - s233564
Mafalda Pires - s232433
Maria Kokali - s232486
Marios Lianos - s233558

April 28, 2024

Contents

1 Introduction

This study is dedicated to applying clustering techniques to identify distinct patterns in the data that correlate with the experiment's phases, seasons, time of the day and the participant roles. The dataset encompasses a range of variables including heart rate metrics (HR), temperature metrics (TEMP), electrodermal activity metrics (EDA-P and EDA-T), and self-rated questionnaire responses. Initially, we carefully examined and prepared the data to make sure it was ready for further analysis. We employed dimensionality reduction methods such as Principal Component Analysis (PCA) and Independent Component Analysis (ICA) to enhance the clarity and interpretability of the data before applying clustering methods - Kmeans, Hierarchical clustering, and Gaussian Mixture Models (GMM). Ultimately, the study draws conclusions highlighting the significant findings and their implications for understanding the metrics from the biosensor and emotional-physiological dynamics within these contexts.

Research questions

- How do the self-rated questionnaire responses vary amongst phases? Are there significant differences between features extracted from the signals relatively to the different phases?
- How does the season (winter or fall) influence the metrics recorded from biosensors, and does each season exhibit distinct characteristics?
- How do the physiological features vary among morning and evening sessions in the fall? Are there any significant differences in the features relatively to the two sessions?
- How do the self-rated questionnaire responses vary among the participant roles in each phase?

2 Data Description - Data Cleaning

The EmoPairCompete dataset comprises physiological data and self-rated emotional responses from questionnaires from 26 participants. The dataset consists of 312 observations and 67 features, which are mostly numerical variables and some of them categorical, and overall there are only 9 missing values. The features regarding the physiological data are numerical variables, while the features Cohort, Round, Phase and the features regarding the questionnaires are considered categorical.

In order to ensure the integrity and maintain consistency across different phases of the experiment, missing values were addressed using two imputation techniques. First, missing values within categorical features were filled using the most frequent value observed within the same experimental phase. For continuous variables, missing entries were imputed using the K-Nearest Neighbors (KNN) algorithm (5 neighbors), which considers the similarity of instances within the same phase to predict the missing values.

Furthermore, besides using the whole dataset, our analysis utilizes several derived datasets from physiological measurements: heart rate features, temperature features, phasic electroderemal activity features, and tonic electroderemal activity features. Each dataset contains features such as mean, standard deviation, minimum, maximum, kurtosis, skewness, slopes, median, area under the curve (AUC), and specific features for EDA such as response time, recovery time, and peaks. Also, in some cases, we used another dataset derived from the features regarding the questionnaires.

Before applying PCA, ICA, and the clustering methods, we standardized all our datasets, which ensures that each feature contributes equally to the analysis, removing biases due to scale differences and enhancing the interpretability and effectiveness of the statistical techniques employed.

3 Methodology

PCA - Principal Component Analysis

We performed PCA in order to represent our dataset in a lower dimension subspace, since we had a lot of variables to consider. PCA transforms our original set of data in a new set, which is a linear combination of the original variables. We chose to kept the principal components (eigenvectors) with corresponding eigenvalues greater than one for the correlation matrix (Kaiser criterion). The components are orthogonal (uncorrelated) and are ordered by the amount of variance they explain in the data.

ICA - Independent Component Analysis

We performed ICA for the same reason as PCA was performed, but under a different assumption. ICA components are also a linear combination of the original variables, but the components are assumed to be statistically independent (in case it is standardized) rather than uncorrelated. ICA aims that the projected data looks as far from Gaussian as possible ([1]).

Clustering methods

In this analysis, we explore the structure and distribution of our dataset using three clustering techniques: K-Means, Hierarchical Clustering, and Gaussian Mixture Models (GMM).

- **K-Means** clustering is a straightforward method that groups data into K separate clusters by minimizing the distance between data points and their respective cluster centroids. Initially, K random points are chosen as cluster centers, and all data points are assigned to the nearest centroid. The centroids are then recalculated based on the mean of the points in each cluster. This process repeats until cluster assignments stabilize. While efficient for large datasets, K-Means requires specifying the number of clusters beforehand and is sensitive to the initial selection of centroids.

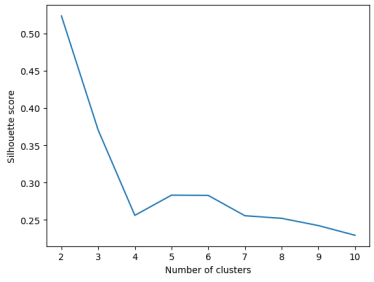
- **Hierarchical clustering** creates a tree-like structure of data, grouping them into clusters step by step without needing a pre-specified number of clusters. This method starts with each data point as its own cluster and progressively merges them based on their closeness, which can be measured in various ways, such as the shortest distance between clusters. The result is often visualized as a dendrogram, illustrating the series of merges and the overall data structure. We chose to use the euclidean distance and the Ward variance minimization algorithm.
- **Gaussian Mixture Models (GMM)** approach clusters by assuming that data points are drawn from several Gaussian distributions, each defined by a mean and variance. Using the Expectation-Maximization algorithm, GMM estimates the parameters of these distributions, thereby accommodating clusters of different shapes and sizes.

4 Clustering conclusions

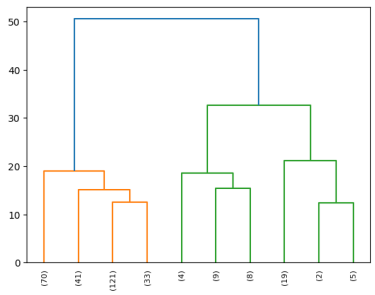
4.1 Phases

For the research question regarding differences in phases, we started by considering the dataset including only the questionnaires responses. It consists of 12 variables. After applying PCA, we get 4 components. In Table 3 we present the component loadings. From its analysis, we can see the first factor has high positive loagings for active, attentive, determined, alert, which could represent a factor related to focus or mental activity. The second component is dominated by frustrated and upset, which captures distress, which contrasts with inspired and determined. The third components is somehow related to fear/anxiety (negative effect of frustration. The forth component is dominated by hostile and ashamed/nervous have negative values. ICA approach was also conducted, but the results were not improved in regards to PCA, so it will be omitted from the report.

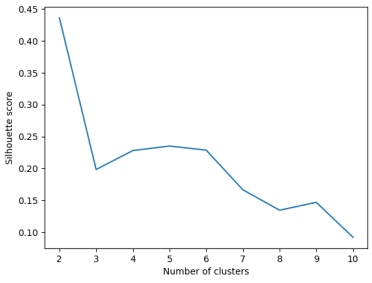
Considering the 4 components we applied the 3 clustering methods cited before.



(a) Silhouette score for K-Means



(b) Dendogram



(c) Silhouette score for GMM

Figure 1

According to Figure 1a, we can see the optimal choice is $K = 2$, which is the same for hierarchical clustering, as presented in Figure 1b, as well as for GMM (1c). In Hierarchical clustering, we defined the number of leaves to 10.

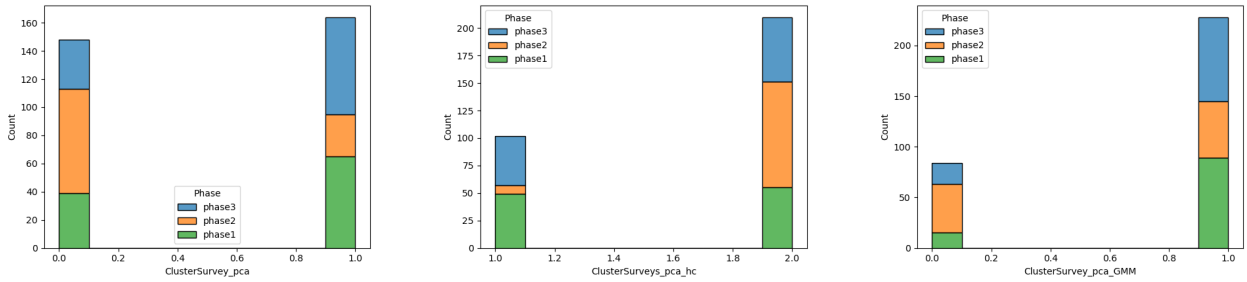


Figure 2: Clustering results: K-Means, Hierarchical and GMM, respectively

According to the results presented in Figure 2, we can see in all some different pattern in phase 2 and the other 2 phases, which have similar patterns. Our initial analysis, showed that phase 1 and 3 were similar, as shown in Figure 4. The number of observations belonging to each cluster are presented in Tabels 4, 5 and 6. Let's break down the results.

	Cluster 1	Cluster 2
Frustrated	1.303922	2.371429
upset	1.117647	1.457143
hostile	1.000000	1.085714
alert	1.411765	2.571429
ashamed	1.058824	1.300000
inspired	1.421569	2.542857
nervous	1.264706	1.438095
attentive	1.774510	3.271429
afraid	1.000000	1.076190
active	1.382353	3.119048
determined	1.941176	3.376190

Figure 3: Average responses - HC

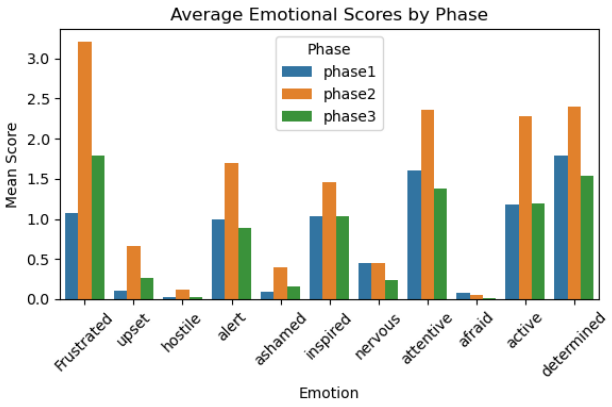


Figure 4: Mean response by question and phase

The results for the other clustering methods are similar and will be omitted. We can conclude from analysis of Table 3 that phase 2 is characterized by an overall of higher questionnaire response values for Frustrated, alert, inspired, attentive, active and determined, whereas the other 2 phases have lower values, but are still spread over the 2 clusters (the values are in the original scale, not standardised to facilitate visualization). This suggests that individuals in Cluster 1 are experiencing these emotions more intensely or frequently than those in Cluster 0.

We can conclude from this analysis, that phase 2 is indeed a phase where participants report higher engagement or activation, which corresponds to periods of higher stress, motivation, or challenge, like the puzzling.

It is also curious that emotions active and determined, where Cluster 1 scores substantially higher, suggest that phase 2 involves activities requiring higher energy and commitment, or that this cluster represents individuals who are in a more engaged state.

4.2 Winter - Fall

In this part of the analysis, we concentrated on different features related to heart rate, temperature, and electrodermal activity. After simplifying our data with the dimensionality reduction techniques mentioned earlier, we applied three clustering methods: K-Means, Hierarchical Clustering, and Gaussian Mixture Models (GMM). Our goal was to explore the clusters formed from these biosensor measurements to determine which participants were involved in the puzzle-solving activities during the winter or fall seasons. However, due to a noticeable difference in the number of participants from Fall (216) compared to Winter (96), we focused on identifying specific characteristics of each category unique patterns.

Optimal number of clusters - PCA

Dataset	HR	TEMP	EDA-P	EDA-T
K-Means Clustering	2	3	2	2
Hierarchical Clustering	3	3	2	2
Gaussian Mixture Clustering	2	4	2	2

Table 1: Optimal number of clusters - PCA

Optimal number of clusters - ICA

Dataset	HR	TEMP	EDA-P	EDA-T
K-Means Clustering	2	4	2	3
Hierarchical Clustering	4	4	4	4
Gaussian Mixture Clustering	2	2	2	2

Table 2: Optimal number of clusters - ICA

In Figure 8 we observe that Cluster 1 within the $ClusterTEMP$ grouping consists of individuals who participated in the experiment during the Winter. Their representation in the other two clusters is notably less pronounced (Cluster 1 = 73, Cluster 2 = 13, Cluster 3 = 10). Moreover, the mean temperature of those participants in Cluster 1 is significantly lower than those from Fall in the same cluster. Additionally, the smaller standard deviation of temperature in Cluster 1 suggests a significant consistency in this measurement, underscoring its relevance for distinguishing participants active in the fall season. Check Figure 14.

Those results are consistent in K-Means clustering with an ICA dimensionality reduction method Figure 12.

We noticed that in every clustering method used, whether we analyzed the $EDA-P$ or $EDA-T$ data and used PCA or ICA for dimension reduction, we consistently identified one cluster that included participants from both the $Winter$ and $Fall$ sessions. Interestingly, the other clusters primarily contained participants from the $Fall$ session. This pattern suggests a unique characteristic or trend among the $Fall$ participants.

After further analysis we conclude that $EDA - T - Mean$, $EDA - T - std$ and $EDA - T - Kurtosis$ are much higher for those participants while $EDA - P - Mean$ and $EDA - P - std$ are higher $EDA - P - Kurtosis$ is lower. Refer to Figure 13 and 14

When we looked at clustering based on heart rate (HR), we found that the clusters were largely shaped by the $HR - Skew$ metrics of the participants (check Figure 15 and 17). Specifically, clusters with higher $HR - Skew$ values predominantly included participants who were part of the Fall session. This suggests a noticeable characteristic related to the heart rate behavior among Fall participants.

4.3 Morning - Evening

For the research question regarding Morning and Evening differences, we worked on the whole dataset and the datasets related to heart rate, temperature and electrodermal activity, but we kept only the observations with Cohorts D3, D4, D5 and D6, which determine only the morning and evening sessions in the fall. We conducted, again, the dimensionality reduction techniques PCA and ICA for all three clustering methods, reduced the complexity of our datasets, and, then, we tried to find the optimal number of clusters for each method, which is number 2 in all cases, after considering the silhouette scores, the scales in our plots and the nature of our research question, which is about identifying the morning and evening sessions.

Since all methods gave us almost similar results, we are going to present here only the insights we got on the whole preprocessed dataset after applying PCA.

The principal component analysis on the whole dataset revealed that 16 principal components explain approximately 81.24% of the variance and can effectively describe the data.

In Figure 18, we can observe the silhouette scores for Kmeans, the dendrogram for Hierarchical clustering and the silhouette scores for GMM clustering. All three plots indicate that the optimal number of clusters is 2.

Now, the results from the three clustering methods:

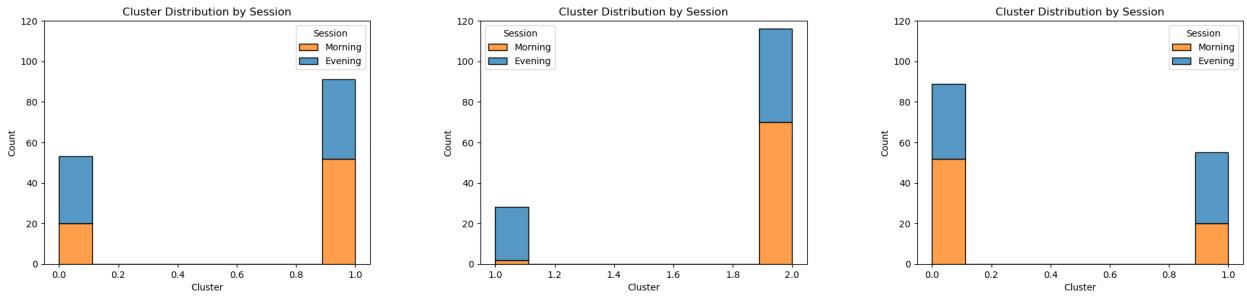


Figure 5: Clustering results: K-Means, Hierarchical and GMM, respectively

- **K-Means:** In Table 13 and in Figure 5, we can observe an overview of the distribution of data points within two clusters, Cluster 0 and Cluster 1. For Cluster 0, the data points are further categorized into two classes, Morning and Evening, with counts of

20 and 33, respectively. Similarly, Cluster 1 comprises data points classified into the same two classes, but with a higher concentration: 52 data points belong to Morning, and 39 to Evening. Cluster 0 has a total of 53 data points, while Cluster 1 has a larger aggregation of 91 data points. Cluster 1 is particularly dominant in class Morning. On the other hand, Cluster 0 shows a relative preference for class Evening.

- **Hierarchical Clustering:** In Table 14 and in Figure 5, we can observe that we have two distinct clusters. Cluster 1 predominantly exhibits characteristics associated with Class Evening, with 26 units compared to only 2 units in Class Morning, suggesting a specific, dominant trait. In contrast, Cluster 2 is more balanced, with 70 units in Class Morning and 46 in Class Evening, indicating a broader set of characteristics. Cluster 1's focus on Evening might imply a precise area of focus, whereas Cluster 2's diversity suggests more general or varied conditions.

- **GMM clustering:** In Table 15 and in Figure 5, we can notice two clusters, each comprising the classes Morning and Evening, too. Cluster 0 contains a total of 89 data points, with a relatively balanced distribution of 52 in the Morning and 37 in the Evening. This suggests a more uniform behavior spread throughout the day. On the other hand, Cluster 1, while smaller with a total of 55 points, presents a split with 20 in the Morning and 35 in the Evening, hinting at a slightly greater emphasis on evening.

Furthermore, based on the above clustering results, we calculated the mean of four physiological features, (heart rate, temperature, phasic and tonic electrodermal activity), for each cluster in every method.

In Table 16, the comparative analysis of physiological features by cluster using KMeans reveals that Cluster 0 displays a lower average heart rate (79.94) than Cluster 1 (82.37), suggesting less physiological arousal. Additionally, Cluster 0 records higher temperatures (33.07 vs. 31.93) and increased electrodermal activity, both tonic (4.71 vs. 1.30) and phasic (0.30 vs. 0.05). These findings indicate that Cluster 0 might experience more pronounced stress responses, possibly due to enhanced sensitivity or more intense reactions to environmental stimuli, compared to the relatively calmer responses observed in Cluster 1.

In Table 17 and 18, after applying Hierarchical clustering and GMM clustering, respectively, we observe similar average values of the physiological features and almost identical patterns as in the K-Means clustering.

Overall, the application of the three clustering methods on our dataset created two clusters, with similar physiological characteristics in all three cases.

4.4 Participant roles

In this study, we sought to determine if distinct emotional profiles emerged for the participant roles during each phase of the activity. We separated the data by Phase, applied PCA and ICA to simplify the complex data and used the clustering methods to group similar responses, supplementing our analysis with ANOVA tests to assess the significance of emotional differences across clusters. All methods gave us almost similar results, thus we

are going to present here the insights we got with K-Means on the dataset that includes only the questionnaires responses, after applying PCA.

Results and Insights

Phase 1

- Clustering yielded two primary groups with varying emotional responses. A noteworthy finding was a significant difference in feelings of frustration (p-value: 0.0165) and activity (p-value: 0.0232), suggesting specific emotional states correspond to physiological changes.
- The clusters consisted of 8 puzzlers and 4 instructors in one, and 44 puzzlers and 48 instructors in another, indicating a balanced distribution of roles within clusters.

As we can see from Figure 6 in the first plot, we have more or less the same distribution in the two clusters. However, in cluster 0 we can see that there are lower emotions average than the cluster 1, and specifically they feel less upset, less active based on the Table 19.

Phase 2

- The silhouette analysis suggested a well-defined clustering with two groups (highest silhouette score: 0.72). There are no significant differences between the emotions in this phase.
- The clusters consisted of 5 puzzlers and 1 instructors in one, and 47 puzzlers and 51 instructors in another.

From Figure 6 in the second plot, we observed that there are more puzzlers in the first cluster, and from the Table 20 we can see that in the first cluster people are more upset, alert and nervous compared to those who are in the second cluster. In the second cluster the distribution of the puzzlers and instructors is almost the same.

Phase 3

- The optimal clustering indicated three primary clusters. 'Frustrated' and 'Active' were significantly different across clusters (p-value: 0.0165, p-value: 0.0232), suggesting a divergence in attention levels or engagement during this phase.
- The clusters consisted of 12 puzzlers and 10 instructors in one, 35 puzzlers and 39 instructors in another and 5 puzzlers and 3 instructors in the last cluster.

As we can see from Figure 6 in the third plot, we observed that the distribution of the puzzlers and instructors are almost the same regarding the three clusters. However, in the second cluster we can see that there are more observations, which indicates that they feel more frustrated, upset, attentive, active and determined based on the table 21. Thus we can conclude that in phase 3 there are much more people that have high levels (more intense) of emotions.

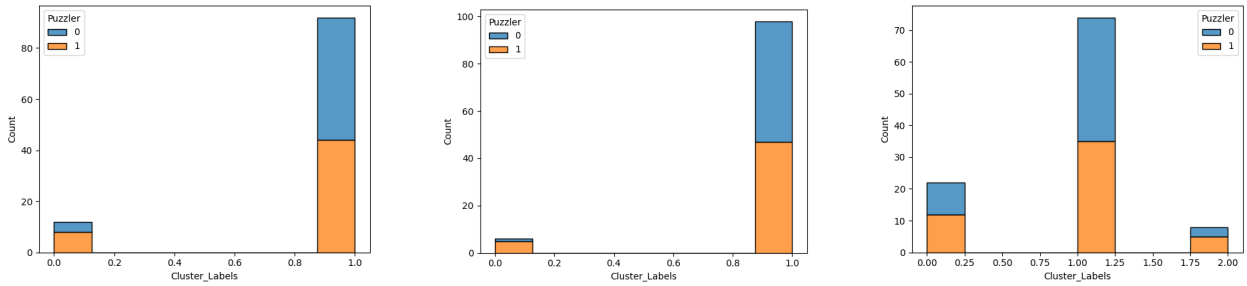


Figure 6: Clustering results: The blue colour indicates the instructors and the orange colour indicated the puzzlers

References

[1] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, The elements of statistical learning: data mining, inference, and prediction, vol. 2. Springer, 2009. Page I of VIII

Appendix

	0	1	2	3
Frustrated	0.122763	0.485767	-0.352968	0.073508
upset	0.118202	0.524248	-0.270612	0.134755
hostile	0.092716	0.379531	0.317482	0.506661
alert	0.414189	-0.012556	-0.081051	-0.145628
ashamed	0.104013	0.404447	-0.127497	-0.543808
inspired	0.363814	-0.267608	-0.099340	0.084223
nervous	0.177158	0.123313	0.480876	-0.586191
attentive	0.444138	-0.075344	0.008031	0.019628
afraid	0.088596	0.231421	0.660707	0.173681
active	0.470141	-0.057760	-0.017491	0.104880
determined	0.436008	-0.178959	-0.023428	0.102628

Table 3: Principal Component Loadings

Cluster	Phase	Observations
0	1	39
	2	74
	3	35
1	1	65
	2	30
	3	69

Table 4: Number of observations per cluster for K-Means

Cluster	Phase	Observations
0	1	49
	2	8
	3	45
1	1	55
	2	96
	3	59

Table 5: Number of observations per cluster for Hierarchical Clustering

Cluster	Phase	Observations
0	1	15
	2	48
	3	21
1	1	89
	2	56
	3	83

Table 6: Number of observations per cluster for GMM

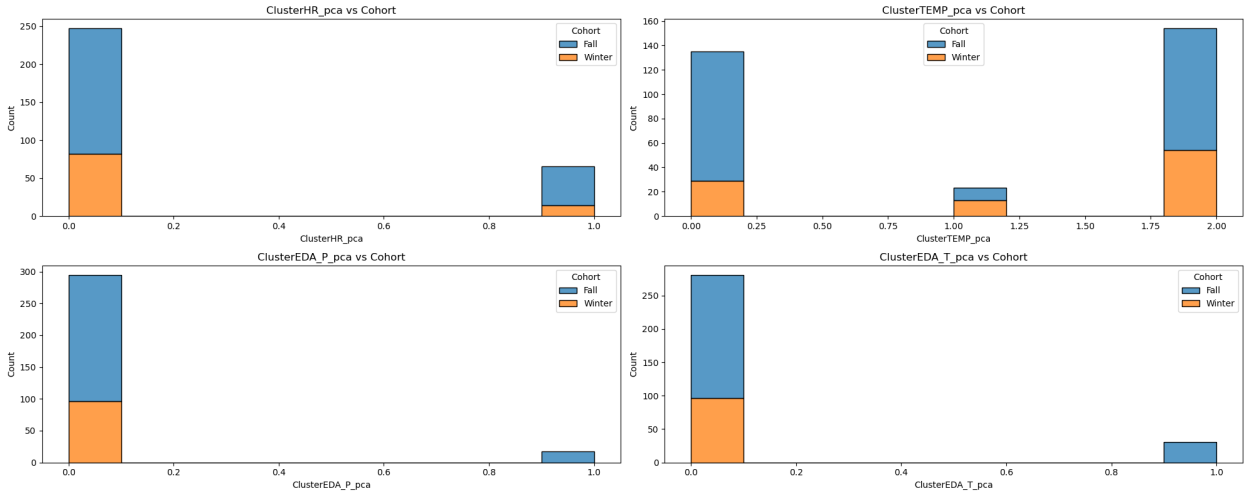


Figure 7: Distribution of 'Season' in K-Means clusters with PCA

Dataset	HR	TEMP	EDA-P	EDA-T
Cluster 1	65	134	295	281
Cluster 2	247	153	17	31
Cluster 3	-	25	-	-

Table 7: Sum of individuals on each cluster

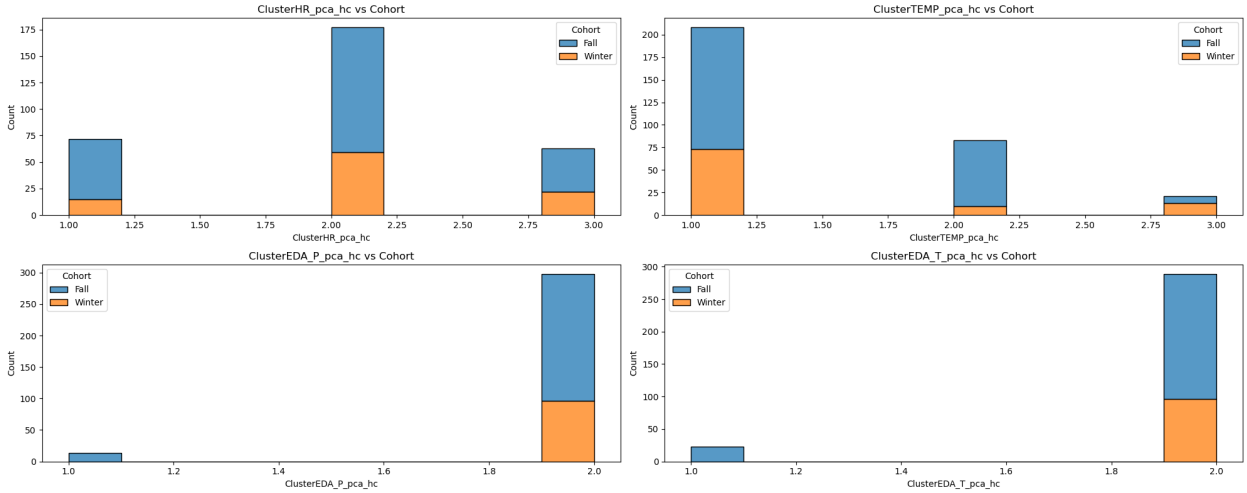


Figure 8: Distribution of 'Season' in Hierarchical clusters with PCA

Dataset	HR	TEMP	EDA-P	EDA-T
Cluster 1	72	208	14	23
Cluster 2	177	83	298	289
Cluster 3	63	21	-	-

Table 8: Sum of individuals on each cluster for K-Means PCA

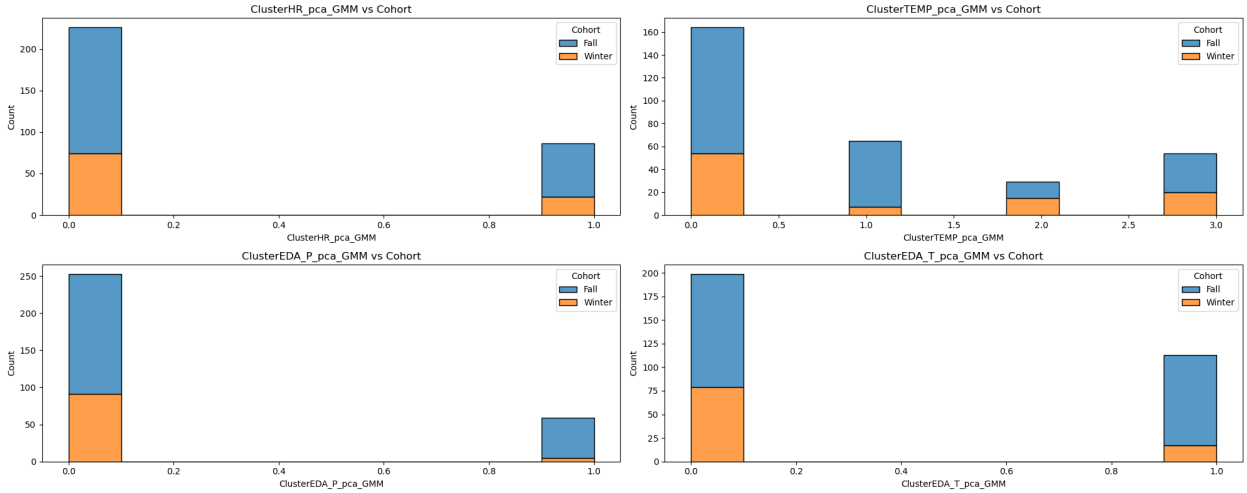


Figure 9: Distribution of 'Season' in GMM clusters with PCA

Dataset	HR	TEMP	EDA-P	EDA-T
Cluster 1	226	54	253	111
Cluster 2	86	164	59	201
Cluster 3	-	65	-	-
Cluster 4	-	29	-	-

Table 9: Sum of individuals on each cluster for K-Means PCA

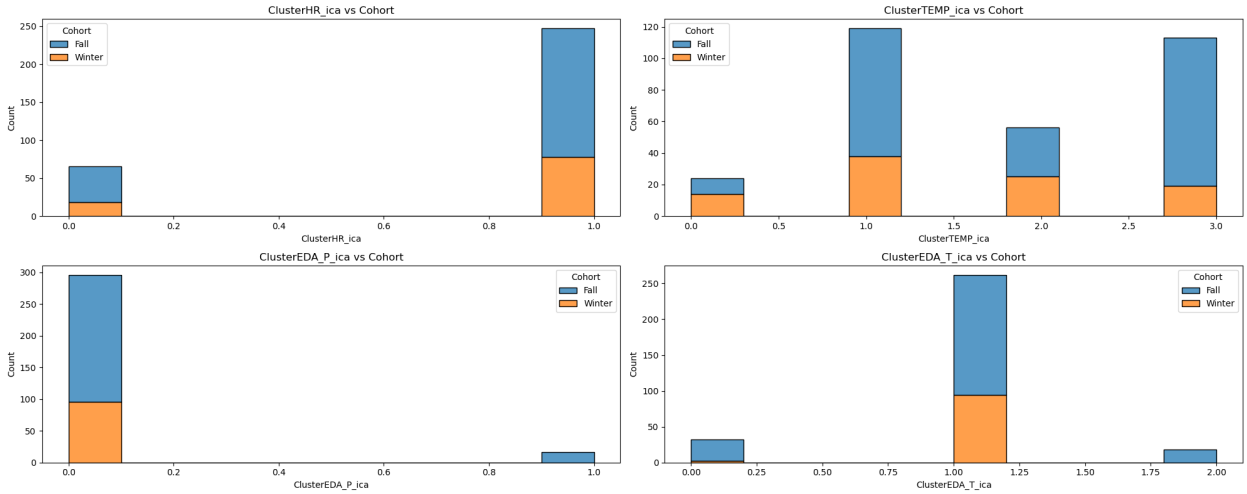


Figure 10: Distribution of 'Season' in K-Means clusters with ICA

Dataset	HR	TEMP	EDA-P	EDA-T
Cluster 1	246	117	16	262
Cluster 2	66	56	296	32
Cluster 3	-	25	-	18
Cluster 4	-	114	-	-

Table 10: Sum of individuals on each cluster for K-Means PCA

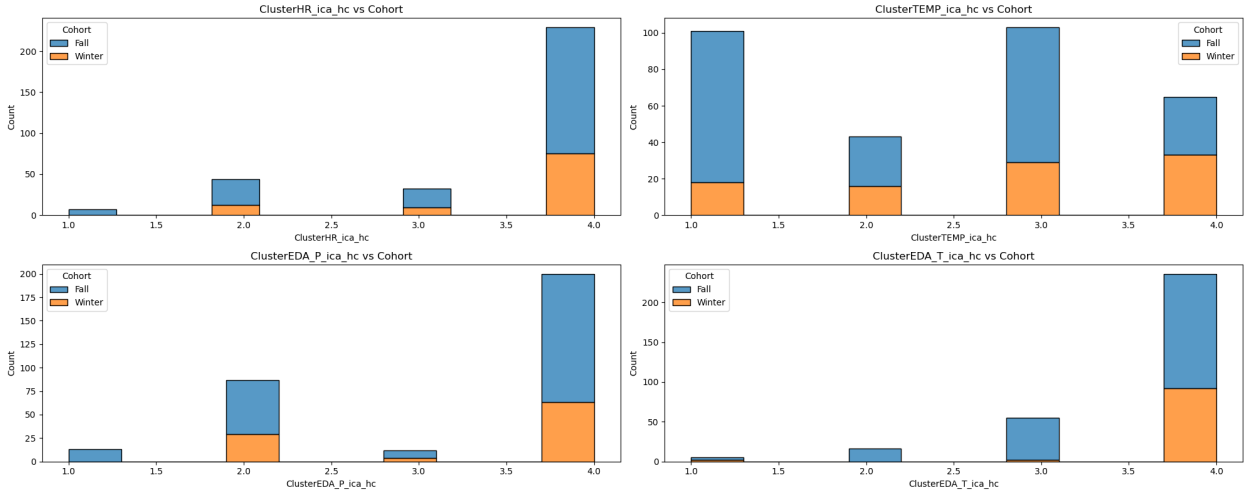


Figure 11: Distribution of 'Season' in Hierarchical clusters with ICA

Dataset	HR	TEMP	EDA-P	EDA-T
Cluster 1	7	101	13	5
Cluster 2	44	43	87	16
Cluster 3	32	103	12	55
Cluster 4	229	65	200	236

Table 11: Sum of individuals on each cluster for Hierarchical ICA

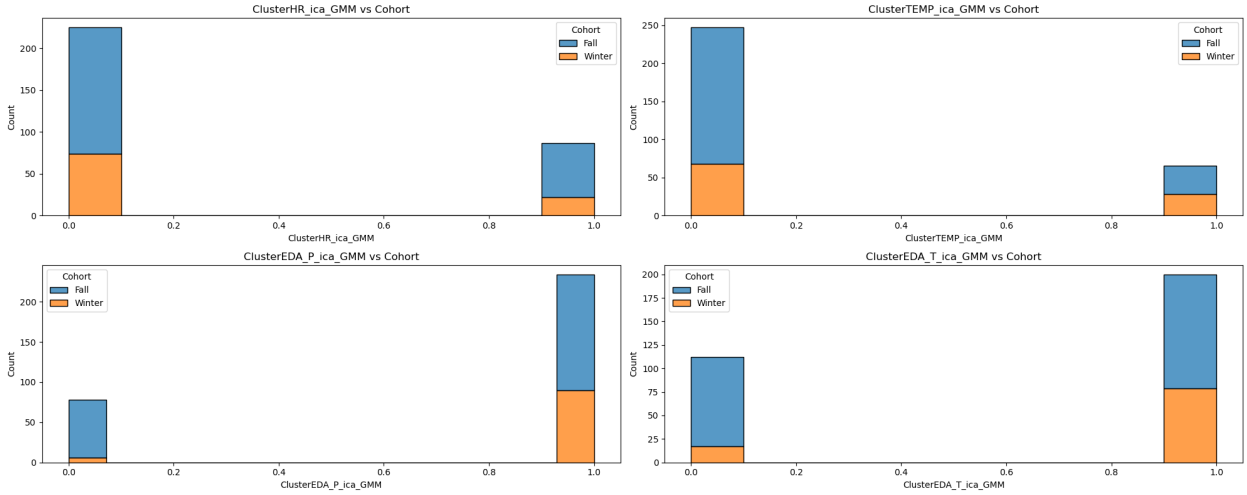


Figure 12: Distribution of 'Season' in GMM clusters with ICA

Dataset	HR	TEMP	EDA-P	EDA-T
Cluster 1	225	247	253	200
Cluster 2	87	65	59	112

Table 12: Sum of individuals on each cluster for GMM ICA

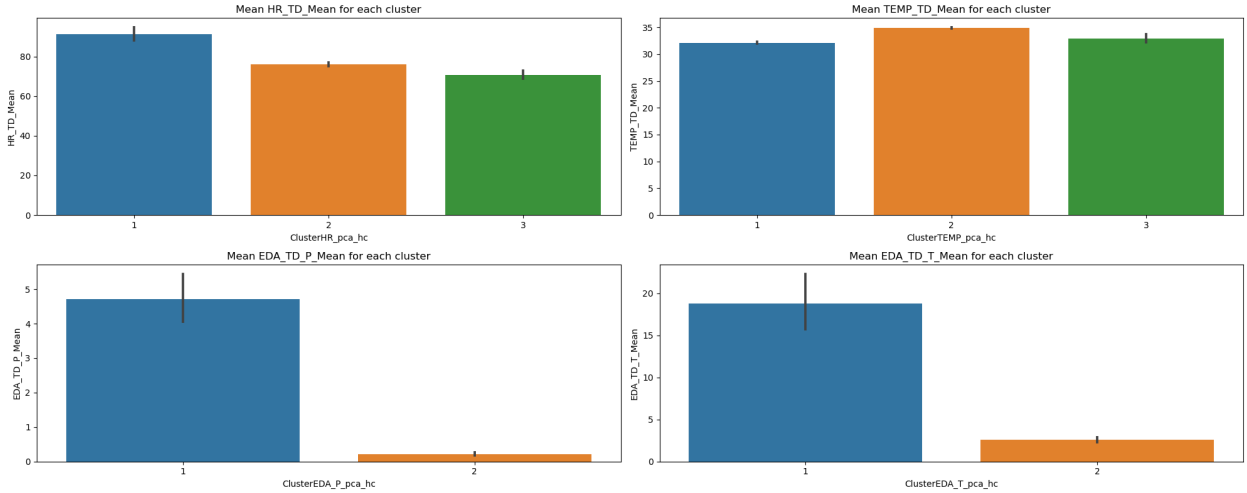


Figure 13: HC Mean with PCA

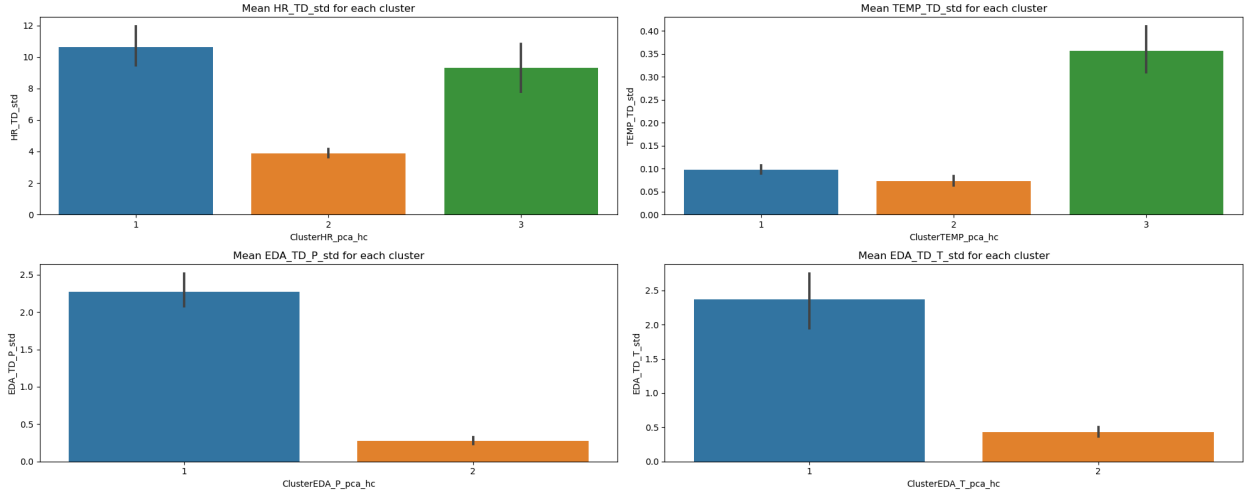


Figure 14: HC TEMP Mean with PCA

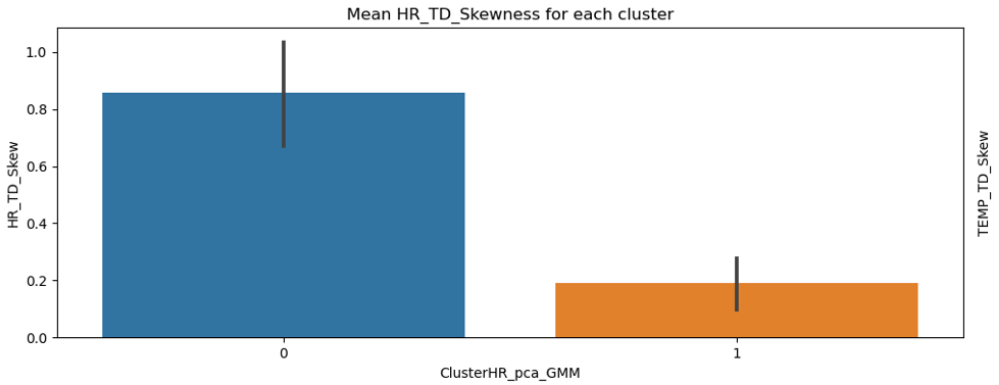


Figure 15: HR Skew GMM with PCA

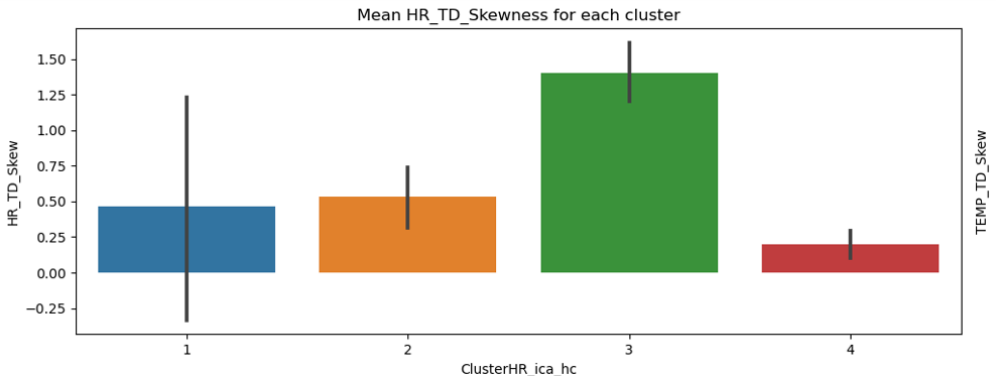


Figure 16: HR Skew HC with ICA

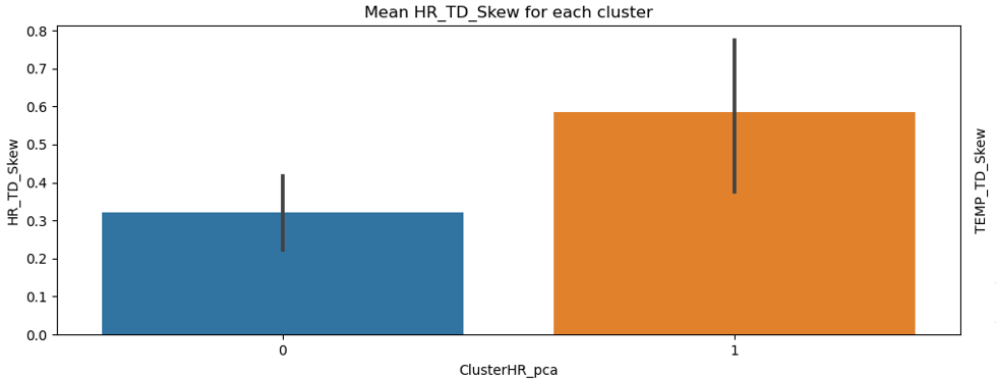


Figure 17: HR SKew K-Means with PCA

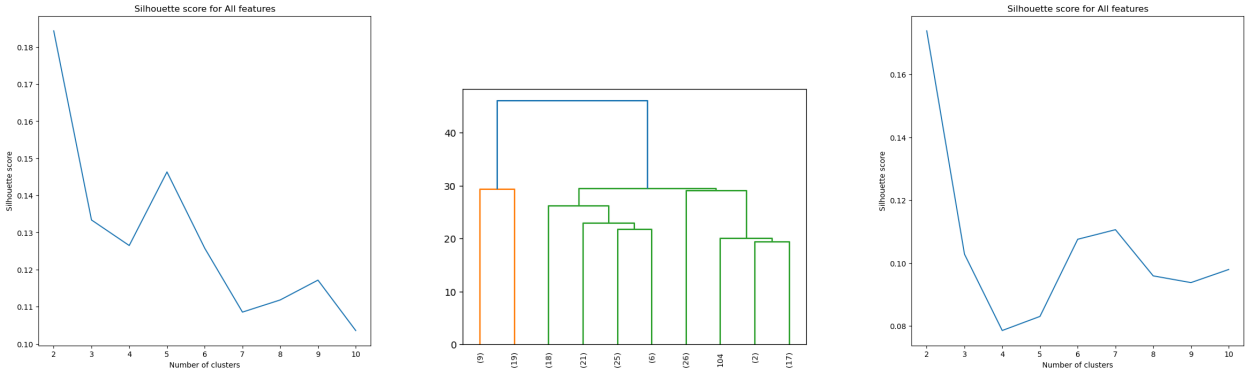


Figure 18: Silhouette scores K-Means, Dendrogram and Silhouette scores GMM, respectively

Cluster	True Class	Count
0	Morning	20
0	Evening	33
1	Morning	52
1	Evening	39

Table 13: Cluster Composition K-Means

Cluster	True Class	Count
1	Morning	2
1	Evening	26
2	Morning	70
2	Evening	46

Table 14: Cluster Composition Hierarchical clustering

Cluster	True Class	Count
0	Morning	52
0	Evening	37
1	Morning	20
1	Evening	35

Table 15: Cluster Composition GMM

Feature	Cluster 0	Cluster 1
HR_TD_Mean	79.94	82.37
TEMP_TD_Mean	33.07	31.93
EDA_TD_T_Mean	4.71	1.30
EDA_TD_P_Mean	0.30	0.05

Table 16: Average values of features for each cluster - KMeans

Feature	Cluster 1	Cluster 2
HR_TD_Mean	79.67	81.91
TEMP_TD_Mean	33.48	32.08
EDA_TD_T_Mean	6.00	1.73
EDA_TD_P_Mean	0.36	0.09

Table 17: Average values of physiological features across clusters - Hierarchical

Feature	Cluster 0	Cluster 1
HR_TD_Mean	76.00	85.39
TEMP_TD_Mean	33.00	31.89
EDA_TD_T_Mean	4.29	1.32
EDA_TD_P_Mean	0.26	0.06

Table 18: Average values of physiological features across clusters - GMM

Emotions	0	1
Frustrated	0.125000	0.236413
upset	0.027778	0.097826
hostile	0.041667	0.010870
alert	0.104167	0.236413
ashamed	0.027778	0.057971
inspired	0.270833	0.258152
nervous	0.125000	0.119565
attentive	0.145833	0.372283
afraid	0.000000	0.007246
active	0.166667	0.315217
determined	0.291667	0.396739

Table 19: Emotions Average Phase 1

Emotions	0	1
Frustrated	0.375000	0.401786
upset	0.388889	0.210884
hostile	0.083333	0.056122
alert	0.458333	0.423469
ashamed	0.111111	0.136054
inspired	0.208333	0.375000
nervous	0.333333	0.219388
attentive	0.666667	0.584184
afraid	0.000000	0.020408
active	0.666667	0.563776
determined	0.583333	0.599490

Table 20: Emotions Average Phase 2

Emotions	0	1	2
Frustrated	0.051136	0.158784	0.140625
upset	0.015152	0.045045	0.000000
hostile	0.000000	0.020270	0.000000
alert	0.170455	0.263514	0.312500
ashamed	0.075758	0.022523	0.000000
inspired	0.215909	0.260135	0.343750
nervous	0.250000	0.195946	0.437500
attentive	0.363636	0.415541	0.375000
afraid	0.000000	0.036036	0.000000
active	0.170455	0.341216	0.218750
determined	0.352273	0.479730	0.406250

Table 21: Emotions Average Phase 3