



**THE AMERICAN
UNIVERSITY IN CAIRO**

الجامعة الأمريكية بالقاهرة

Assessing Neurological States from Physiological Signals

Data Mining Project

Fall 2024

Report Conducted By:

Youssef Nakhla - 900201430

Zeina Kishk- 900211723

Hanya Sheikh - 900212533

Table of Contents

- 1. Introduction**
- 2. Dataset**
- 3. Exploratory Data Analysis**
- 4. Preprocessing**
- 5. Unified Pipeline**
- 6. Modeling**
- 7. Classification Models**
 - 6.1. Support Vector Machine (SVM)
 - 6.2. Multi-Layer Perceptron (MLP)
 - 6.3. K-Nearest Neighbors (KNN)
- 8. Clustering Models**
 - 7.1. K-Means Clustering
 - 7.2. Hierarchical Clustering
 - 7.3. Spectral Clustering
- 9. Conclusion**

Introduction:

Stress is a relevant condition that significantly impacts physical, cognitive, and emotional well-being. Accurate detection and classification of stress levels are critical for understanding its effects on the human body and for developing interventions to mitigate its impact. This project aims to assess neurological states, including physical, emotional, and cognitive stress, using non-invasive physiological signals utilizing the power of machine learning to ease the process of classifying these neurological states. In this project, we will experiment with clustering methods to group similar physiological responses and classification methods to categorize the different neurological states. The goal is to develop an accurate model that can identify and classify stress and relaxation states based on these physiological signals.

Dataset:

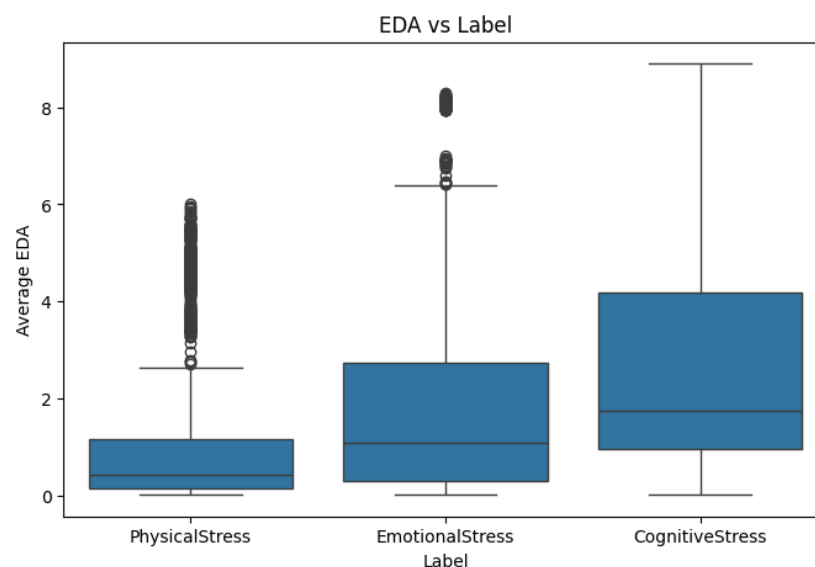
Our dataset is the Non-EEG Dataset for Assessment of Neurological Status on PhysioNet contains physiological data collected from healthy volunteers under various stress conditions. The subjects in this dataset have the demographics of age ranging from 21–28 years with 11 males and 9 females. The dataset used was collected from 20 healthy subjects at the University of Texas at Dallas and includes signals such as electrodermal activity (EDA), heart rate (HR), acceleration, temperature, and arterial oxygen level (SpO2).

These signals were recorded while the subjects went through various stress-inducing tasks, such as physical exercise, cognitive challenges, and emotional stress like watching a horror movie. The dataset consists of 7 stages per subject, including relaxation and different stress states. The data is provided in WFDB format, with separate records for each type of signal and

annotations marking transitions between stages. Cognitive stress refers to stress that primarily affects mental processes like thinking, memory, and decision-making. Physical stress impacts the body directly, often resulting in physiological changes and discomfort. Finally, Emotional stress that affects feelings and emotional well-being, often tied to interpersonal or internal conflicts.

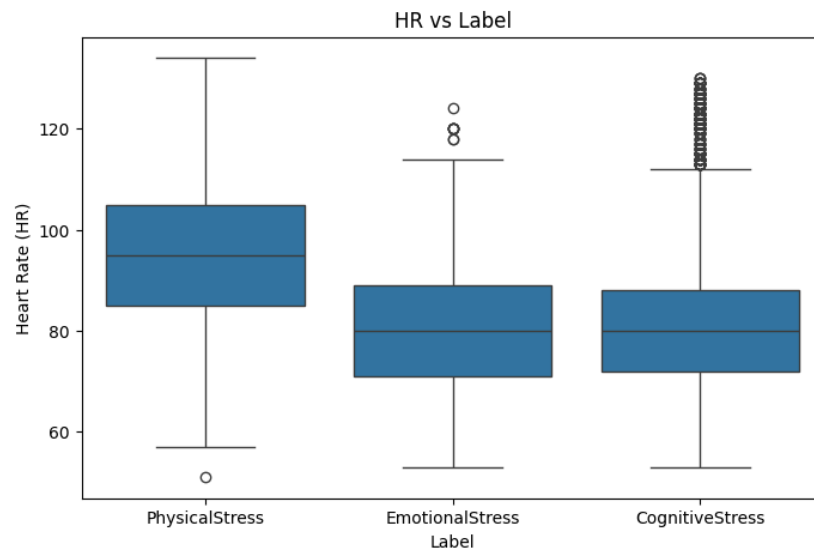
Exploratory Data Analysis:

We wanted to explore the relationships between each variable and the three different types of stress hence we plotted box plots to visualize the effect of each variable on the stress state. Firstly, we plotted the Electrodermal Activity (EDA) with the 3 labels and the results can be seen below. Since our label is discrete and can only take one of four values, the box plot was most appropriate to visualize the relationships. The relationship between EDA and our labels are that the higher the value of EDA the more likely it is that this state correlates with Cognitive stress, and with a decrease in value of EDA, the state would be Emotional Stress and then followed by Physical stress.

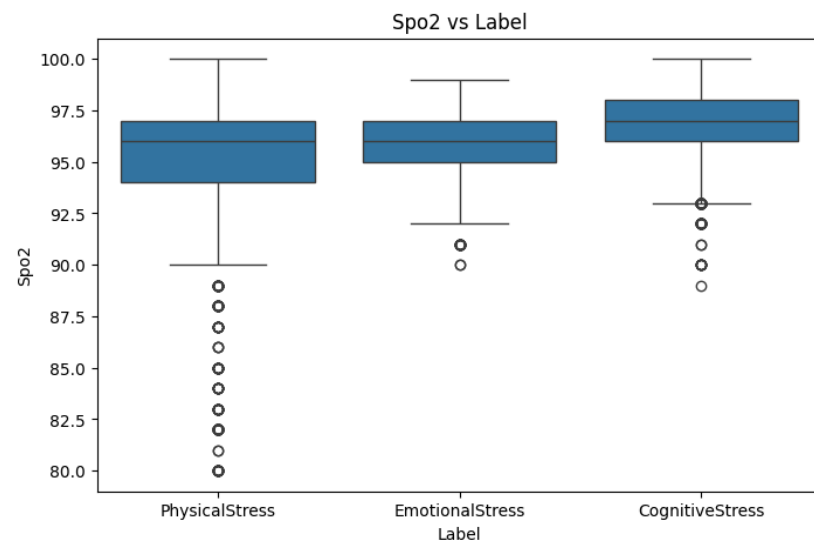


Next we plotted the heart rate with the 3 labels and the results can be seen below;

contrary to the relationship between the EDA and the label, the greater the value of the heart rate the more likely it is to be Physical stress, followed by Emotional and then Cognitive stress.



Lastly, we plotted the oxygen saturation with the three labels and the results can be seen below; the graph shows that the value of SPO2 on the stress state is almost constant for all states with Cognitive stress taking a bit of lead.



Preprocessing:

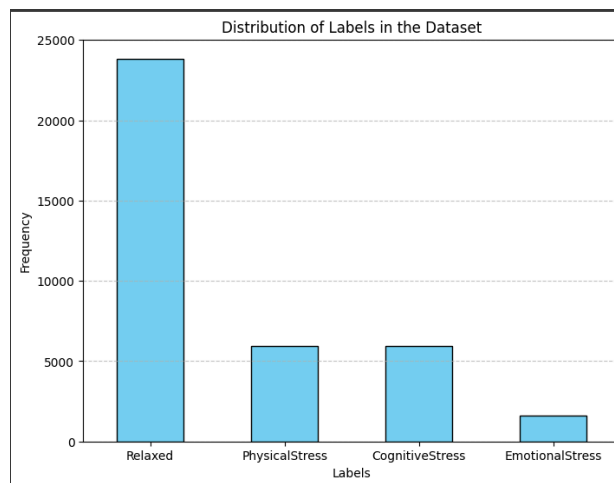
For the preprocessing phase, we began by loading the signals for 20 subjects which included SpO2, heart rate, acceleration, temperature, and electrodermal activity (EDA). The dataset contained seven stages per subject: three relaxation stages and four stress stages (PhysicalStress, EmotionalStress, CognitiveStress, and Mini-emotional Stress). To simplify the analysis, we combined the three relaxation stages into a single category labeled “Relaxed.” We also merged the "Mini-emotional stress" and "Emotional Stress" stages into one category. As a result, we have four states: Relaxed, Cognitive Stress, Emotional Stress, and Physical Stress. The transitions between stages were marked in the data with annotations, which we used to divide the signals into separate segments for each stage. Since the signal lengths varied, we resized all signals to match the minimum available length for each stage, ensuring uniformity across the dataset.

Next, we combined the segmented signals into feature vectors, where each vector represented a subject's physiological responses during a specific stage. Each vector was labeled based on the stage it belonged to, whether it was a relaxation stage or one of the stress stages (Physical, Emotional or Cognitive). The data was split into sub variables that are summarized in the table below.

Variables	Description
ax, az, ay	These represent the acceleration data along the x-axis, y-axis, and z-axis, respectively

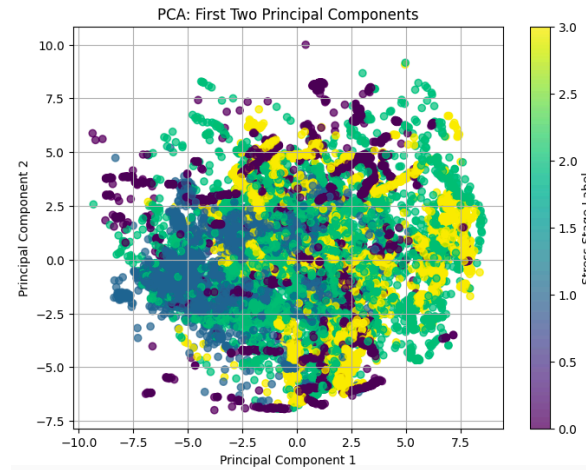
EDA	Electrodermal Activity, measures changes in skin conductance.
temp	temperature data
spo2	oxygen saturation level in the blood
hr	heart rate

Furthermore, the demographic data of each subject was included in the dataset along with numerically encoding the labels such that 0 stands for Relaxed, 1 stands for Physical Stress, 2 stands for Emotional Stress, and 3 stands for Cognitive Stress. In addition, the gender was also numerically encoded such that females took a label of 0 and males took a label of 1.



In the figure above, there is a visible class imbalance. To tackle the issue, we first applied undersampling to the relaxation class by calculating cosine similarity scores and keeping only the most distinct samples. To address the underrepresentation of the CognitiveStress state, we applied the SMOTE technique to generate additional samples. This helped balance the dataset, ensuring that each of the four states had an equal representation of 5,960 samples.

We then applied three different dimensionality reduction methods separately to prepare the data for further analysis. First, we used Principal Component Analysis which reduced the features to 8 components while retaining 95% of the data's variance.



Next, we applied Sequential Feature Selection which iteratively selects the most important features based on model performance. Lastly, we used Recursive Feature Elimination to eliminate less important features, keeping only those that contribute the most to classification accuracy. Each of these dimensionality reduction techniques was tested independently, and we will evaluate their effectiveness by applying them to our classification and clustering models. This will allow us to determine which dimensionality reduction method provides the best performance for our next step.

Unified Pipeline:

To simplify the process of testing different combination dimensionality reduction, feature extraction, and machine learning models, a unified framework was developed that is built on initialized functions where each section of the code was defined separately so that different combinations can be tested with ease, please refer to the code to further examine this framework

Modeling:

We wanted to figure out how to maximize the performance of our models, hence, we tested our models with different data such that one iteration the models one iteration was tested on data only using PCA, another was tested on data only using SFS, etc. This was done to see which combination is the most robust and efficient for classification/clustering. The sections below will dive deeper into the work done.

Classification Models:

We used three different classification methods: Support Vector Machine, Multi-Layer Perceptron, and K-Nearest Neighbors to classify the four states: Relaxed, Cognitive Stress, Emotional Stress, and Physical Stress. We split the data into training, validation, and test sets to assess the models' performance. Since both Recursive Feature Elimination and Sequential Feature Selection resulted in the same reduction of features, we decided to only use PCA and SFS for dimensionality reduction to test our three different models separately. Each model was trained and validated on the reduced datasets, and we compared their performance based on accuracy, F1 score, and confusion matrices.

a) Support Vector Machine

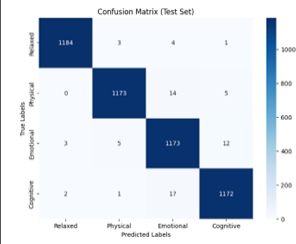
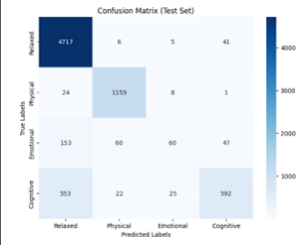
We applied the support vector machine method to classify the four different states. SVM is a powerful supervised learning algorithm that works by finding a hyperplane that best separates the classes in the feature space. We chose SVM for this task because it is well-suited for high-dimensional data like physiological signals. The results of the different dimensionality techniques are summarized in the table below:

	Accuracy	F1 Score	Confusion matrix																									
PCA	0.933	0.933	<table><caption>Confusion Matrix (Test Set)</caption><thead><tr><th></th><th>Relaxed</th><th>Physical</th><th>Emotional</th><th>Cognitive</th></tr></thead><tbody><tr><th>Relaxed</th><td>1110</td><td>4</td><td>25</td><td>18</td></tr><tr><th>Physical</th><td>0</td><td>1168</td><td>23</td><td>1</td></tr><tr><th>Emotional</th><td>20</td><td>45</td><td>1042</td><td>86</td></tr><tr><th>Cognitive</th><td>16</td><td>5</td><td>65</td><td>1106</td></tr></tbody></table>		Relaxed	Physical	Emotional	Cognitive	Relaxed	1110	4	25	18	Physical	0	1168	23	1	Emotional	20	45	1042	86	Cognitive	16	5	65	1106
	Relaxed	Physical	Emotional	Cognitive																								
Relaxed	1110	4	25	18																								
Physical	0	1168	23	1																								
Emotional	20	45	1042	86																								
Cognitive	16	5	65	1106																								
SFS	0.721	0.377	<table><caption>Confusion Matrix (Test Set)</caption><thead><tr><th></th><th>Relaxed</th><th>Physical</th><th>Emotional</th><th>Cognitive</th></tr></thead><tbody><tr><th>Relaxed</th><td>4031</td><td>78</td><td>0</td><td>0</td></tr><tr><th>Physical</th><td>408</td><td>704</td><td>0</td><td>0</td></tr><tr><th>Emotional</th><td>305</td><td>13</td><td>0</td><td>0</td></tr><tr><th>Cognitive</th><td>1123</td><td>69</td><td>0</td><td>0</td></tr></tbody></table>		Relaxed	Physical	Emotional	Cognitive	Relaxed	4031	78	0	0	Physical	408	704	0	0	Emotional	305	13	0	0	Cognitive	1123	69	0	0
	Relaxed	Physical	Emotional	Cognitive																								
Relaxed	4031	78	0	0																								
Physical	408	704	0	0																								
Emotional	305	13	0	0																								
Cognitive	1123	69	0	0																								

b) Multi-Layer Perceptron

For our second approach, we applied a multi-layer perceptron to classify the four different states. MLP is a type of neural network with multiple layers of nodes that helps model complex relationships in the data. It is particularly good for tasks with non-linear decision boundaries, such as classifying physiological signals, which can have complicated patterns so we decided to try it out. The MLP model was trained with the default hidden layer of 100 units and up to 1000 iterations to make sure the model converged. The results of the different dimensionality techniques are summarized in the table below:

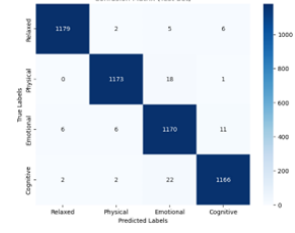
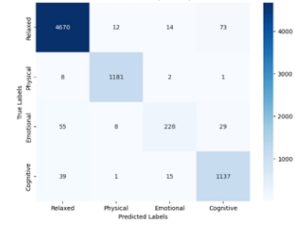
	Accuracy	F1 Score	Confusion matrix
--	----------	----------	------------------

PCA	0.985	0.985	 <p>Confusion Matrix (Test Set)</p> <table><tr><th></th><th>Relaxed</th><th>Physical</th><th>Emotional</th><th>Cognitive</th></tr><tr><th>Relaxed</th><td>1184</td><td>3</td><td>4</td><td>1</td></tr><tr><th>Physical</th><td>0</td><td>1173</td><td>14</td><td>5</td></tr><tr><th>Emotional</th><td>3</td><td>5</td><td>1173</td><td>12</td></tr><tr><th>Cognitive</th><td>2</td><td>1</td><td>17</td><td>1172</td></tr></table>		Relaxed	Physical	Emotional	Cognitive	Relaxed	1184	3	4	1	Physical	0	1173	14	5	Emotional	3	5	1173	12	Cognitive	2	1	17	1172
	Relaxed	Physical	Emotional	Cognitive																								
Relaxed	1184	3	4	1																								
Physical	0	1173	14	5																								
Emotional	3	5	1173	12																								
Cognitive	2	1	17	1172																								
SFS	0.873	0.698	 <p>Confusion Matrix (Test Set)</p> <table><tr><th></th><th>Relaxed</th><th>Physical</th><th>Emotional</th><th>Cognitive</th></tr><tr><th>Relaxed</th><td>4717</td><td>6</td><td>5</td><td>41</td></tr><tr><th>Physical</th><td>24</td><td>1159</td><td>8</td><td>1</td></tr><tr><th>Emotional</th><td>153</td><td>60</td><td>60</td><td>47</td></tr><tr><th>Cognitive</th><td>553</td><td>22</td><td>25</td><td>392</td></tr></table>		Relaxed	Physical	Emotional	Cognitive	Relaxed	4717	6	5	41	Physical	24	1159	8	1	Emotional	153	60	60	47	Cognitive	553	22	25	392
	Relaxed	Physical	Emotional	Cognitive																								
Relaxed	4717	6	5	41																								
Physical	24	1159	8	1																								
Emotional	153	60	60	47																								
Cognitive	553	22	25	392																								

c) K-Nearest Neighbours

For our last approach, We applied the KNN algorithm to classify the four different states. KNN is a distribution-free method because it doesn't assume any specific distribution for the data. It classifies a data point based on the majority class of its nearest neighbors without relying on any predefined model or structure. The model was trained using a default of 5 neighbors to classify the states based on the physiological data. This method we thought would be good for our dataset because the physiological signals may not follow a typical pattern or distribution. The results of the different dimensionality techniques are summarized in the table below:

	Accuracy	F1 Score	Confusion matrix
--	----------	----------	------------------

PCA	0.983	0.983	
SFS	0.965	0.922	

The results from the classification models showed that all three models SVM, MLP, and KNN performed well with PCA dimensionality reduction. MLP and KNN performed better than SVM, with MLP achieving the highest accuracy and F1 score. PCA outperformed SFS in improving model performance, as it resulted in higher accuracy and F1 scores across all models. Overall, PCA proved to be more effective than SFS for this task, enhancing the models' ability to accurately classify the different neurological states with high accuracy.

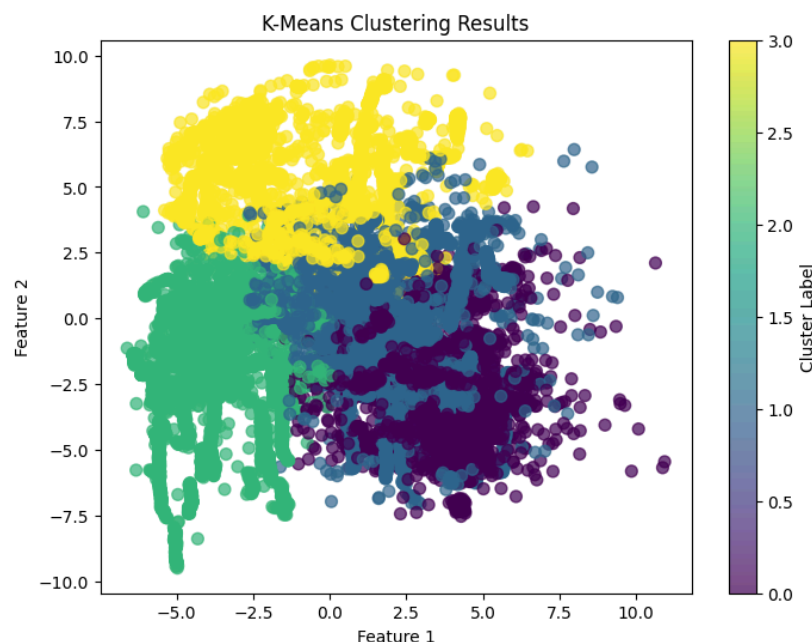
Clustering Models:

In addition to the classification models, we decided to test three clustering models to see how unsupervised learning would do with our data and problem. We tried K-means clustering, Spectral clustering and Hierarchical clustering, and since unsupervised learning does not require the splitting of data to training and testing, we dropped the label from each. In our implementation, we applied the clustering models to group the dataset into four clusters, corresponding to the four stress categories: Relaxed, Physical Stress, Cognitive Stress, and

Emotional Stress. For all three models the silhouette score, f1 score and accuracy were computed to evaluate clustering quality, with higher scores indicating better-defined and more distinct clusters. Before the application of PCA to our data all clustering methods were very dense in terms of time complexity and due to gpu constraints they were unable to run, hence we decided to only test clustering models after the implementation of PCA.

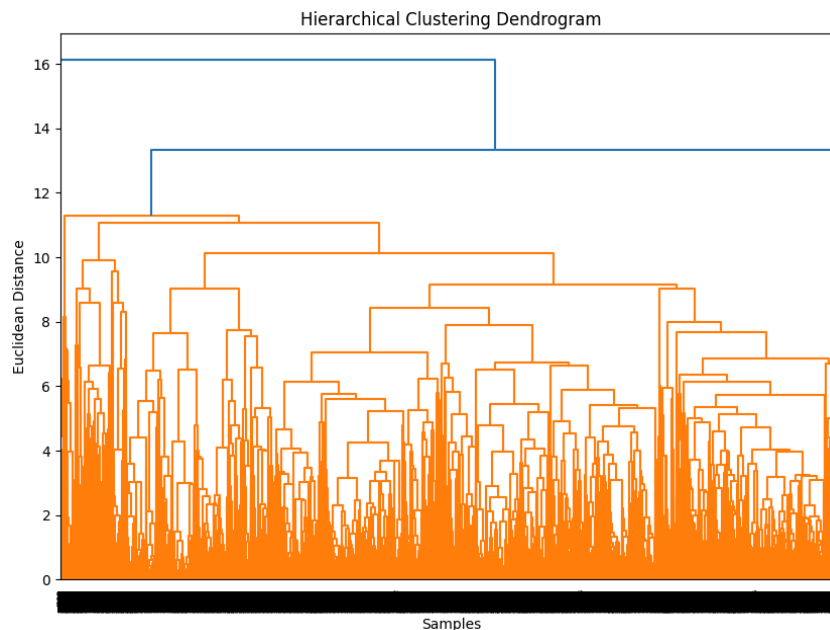
a) K-means clustering

K-Means is an **unsupervised learning algorithm** used for partitioning data into distinct groups (clusters) based on similarity. It iteratively assigns data points to clusters and updates cluster centers to minimize intra-cluster variance. The model was initialized with the k-means++ method to improve cluster center initialization, and the number of iterations was limited to 300 to ensure convergence. After fitting the model to the data, the cluster labels were predicted for all observations. The results of K-means clustering as can be seen below were very poor achieving an accuracy of 0.27.



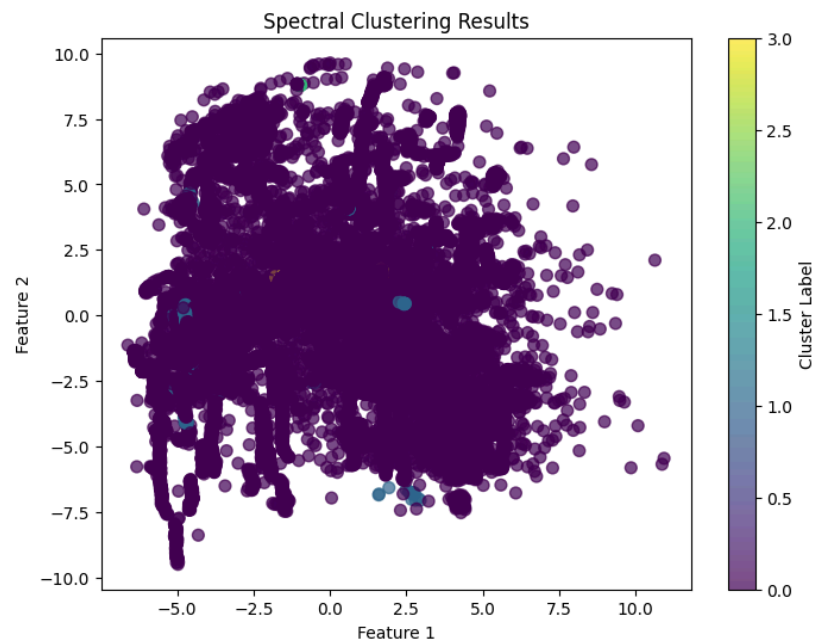
b) Hierarchical clustering

Hierarchical clustering is an unsupervised learning algorithm that builds a hierarchy of clusters by iteratively merging or splitting data points based on similarity. In our implementation, we used agglomerative hierarchical clustering with Ward's linkage to minimize intra-cluster variance during each merging step, ensuring compact and homogeneous clusters. After fitting the model, cluster labels were assigned to all observations. Additionally, a dendrogram was generated to visualize the hierarchical structure of the clusters, providing insight into the relationships and distances between different data points. This approach allowed us to analyze the dataset's natural grouping patterns and evaluate cluster separability. As expected based on the results of the K-means clustering the Hierarchical clustering also performed badly, retaining and accuracy of 0.25 with the dendrogram being shown below.



c) Spectral clustering

Finally, Spectral clustering is an unsupervised learning algorithm that uses the graph-based representation of data to partition it into distinct clusters. The algorithm constructs a similarity graph of the data, where each data point is connected to its nearest neighbors, and uses the graph's spectral properties to find optimal cluster assignments. After fitting the model, cluster labels were assigned to all observations. Spectral clustering is particularly effective for non-convex or complex cluster shapes, providing an alternative approach to identify underlying structures in the dataset. This technique allowed us to explore cluster separability in the context of stress classification. Spectral clustering also performed very badly with an accuracy of 0.15.



The results from the clustering models demonstrated poor performance in accurately separating the four stress categories in this dataset. Firstly, the physiological signals in the dataset exhibit significant overlap in feature space across the different stress categories. Clustering algorithms rely purely on feature similarity to group data points, but the inherent overlap makes it difficult to distinguish between the categories, especially when the stress responses are subtle

or not distinctly separated. Also, clustering methods are sensitive to noise and feature scaling, and while PCA helped reduce dimensionality and improve computational efficiency, it may also have removed subtle variations that were critical for differentiating stress categories.

Conclusion:

In this project, we tested both classification and clustering methods to classify and group the four neurological states: Relaxed, Cognitive Stress, Emotional Stress, and Physical Stress. For the classification, we tried **Support Vector Machine (SVM)**, **Multi-Layer Perceptron (MLP)**, and **K-Nearest Neighbors (KNN)** models. We found that PCA dimensionality reduction worked best for all three models, improving their accuracy and F1 score. MLP and KNN performed the best, with SVM having slightly lower results. We also tried three clustering models: **K-means**, **Hierarchical clustering**, and **Spectral clustering**, to see how unsupervised learning would work with our data. Unfortunately, these models performed poorly, with accuracy ranging from 0.15 to 0.27. This is likely because the physiological signals for different stress states overlapped a lot, making it difficult for the clustering algorithms to separate the states. Although PCA helped reduce the data's complexity and improved computation, it didn't fix the issue of overlapping features for clustering. In conclusion, the classification models worked well, but the clustering models struggled due to the overlap in the data. This shows that while unsupervised learning could be helpful, it might require better ways to separate the features to work effectively in this context.