Classification on Non-EEG Dataset for Assessment of Neurological Status

Yehya Abdelmohsen, Maria Amgad, and Mona Ahmad

Introduction

Stress can be the result of several aspects of daily life. It is recognized by the brain as the neurological signals change to indicate different types of stress such as physical, emotional, and cognitive stress. However, there are other non-brain physiological signals that can be used to determine the neurological state. These non-EEG physiological signals include electrodermal activity (EDA), which represents the electrical characteristics of the skin, temperature, acceleration, heart rate (HR), and arterial oxygen level (SpO2). These signals are measured by various biosensors such as the accelerometer and the EDA biosensor. These signals will be used to identify four main phases, which are relaxation, physical stress, emotional stress, and cognitive stress.

Problem Definition

The non-EEG physiological signals used in the project were obtained from the Quality of Life Laboratory at University of Texas at Dallas [1]. The data were collected from 20 healthy people. The purpose is to develop a model that takes the input signals and classifies it to its appropriate phase, where the four phases are relaxation, emotional stress, cognitive stress, and physical stress. Number of signals input as well as their description are provided in the data description section.

Data Description

There are twenty subjects in the collected data. For each subject, the temperature, EDA signals, acceleration, heart rate, and SpO2 were measured. The accelerometer was the device used to measure the acceleration of the subject; it provides x, y, and z coordinates showing the direction and position of human movement/acceleration. Therefore, there are three signals collected for the accelerator measurements, which are known as ax, ay, and az. In total, each subject has seven signals: three accelerometer signals, and the remaining are the temperature, EDA signals, heart rate, and SpO2 signals.

The data for each subject is provided in three files; two of them are WFDB format files, where the first file includes the accelerometer, temperature, and EDA signals while the other includes SpO2 and heart rate signals. For the first file, eight samples were taken per second while in the second file, only one sample was recorded per second. The last file is an annotation file that shows the time and location where the signals change from one phase to the next and the name of each phase. There is also a csv file and a header file that include some information about the subjects as their age, gender, height, and weight. However, the header and csv files were not included when developing the model. Figure 1 and figure 2 show an illustration of all of the information of subject 1.

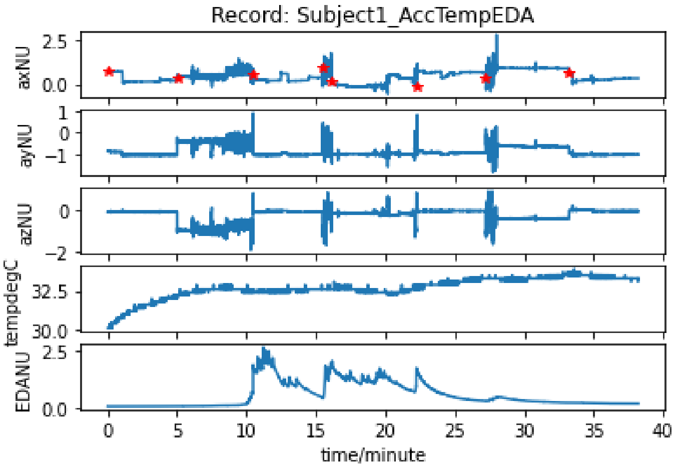

Figure 1. Accelerometer, Temperature, and EDA signals

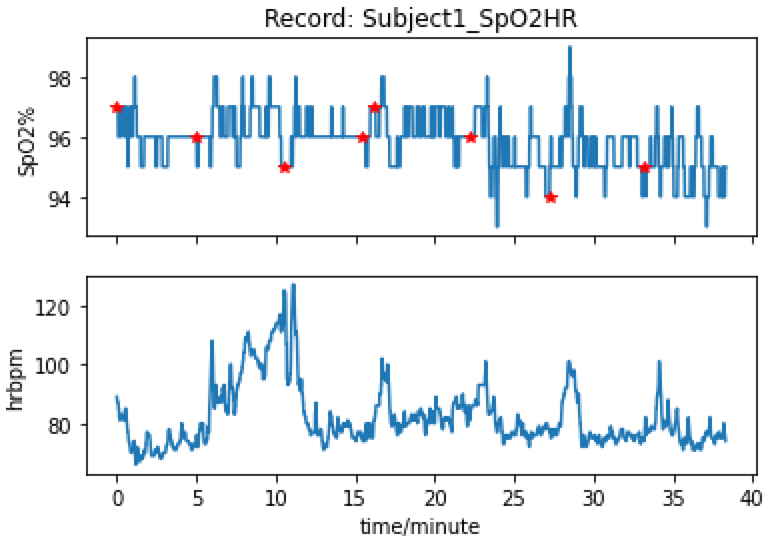

Figure 2. SpO2 and HR signals

All of the twenty healthy people underwent the same experimental procedures, which aimed to let the subjects go through seven stages. Each subject relaxes for five minutes, then is asked to do a physical activity (walking/jogging), then are given another five minutes to relax [1]. Afterwards, they underwent the mini-emotional stress stage, which lasted for 40 seconds, followed by the three-minute cognitive stress which involved backward counting. After that, the person was given five minutes to relax, then he went under emotional stress due to a horror movie for five minutes, and the experiment ended with five minutes of final relaxation.

While the main four stages are relaxation, emotional stress, cognitive stress, and physical stress, it is important to recognize that the input signals include four parts in the relaxation phase, two parts in the emotional stress phase, one in cognitive stress phase, and one in physical stress. For each of the seven experimental stages, the non-EEG physiological signals that were previously measured were recorded.

Methodology

1. Data Manipulation

A class called subject was created to store the seven signals of each subject. Since different subjects have different signal sizes, the signals had to be manipulated to ensure consistency when creating the feature vector, which will be provided to the model during the training phase

The cutoff values provided in the annotation file of each subject were used to identify the separate stages within the signal for all of the 20 subjects. For each of the four stages, the minimum length of stage was determined. Since accelerometer, temperature and EDA signals, which are known as group 1, were recorded at a rate of 8 samples/second while heart rate and SpO2, which are known as group 2, were obtained at a rate of 1 sample/second, the minimum value for each phase was obtained for each group. Subsequently, the signals were resized according to the minimum value of each phase and any extra values more than the minimum length were discarded. This step ensured that signals of different subjects maintained the same size.

2. Feature Vector

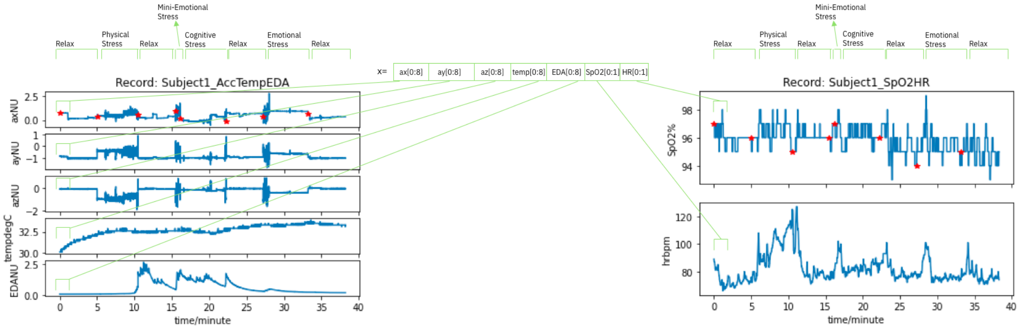

Figure 3. Feature Vector

After splitting the signals into its stages using the information provided in the annotation files, for each subject, each stage was divided into multiple feature vectors. For group 1 signals, which are recorded at a rate of 8 samples/second, the size of the data that will be stored in the feature vector is defined by eta, which is a tunable parameter, with the constraint that the integer must be divisible by 8. This value determines the size of the split and value that will be used in each feature vector for group 1 signals. For group 2 signals, the amount of data for each signal that will be stored in the feature vector is calculated by dividing eta over 8, as signals in the second group are recorded at a rate of one sample/second. An illustration of the feature vector is shown in figure 3. This process is repeated over all signals, producing 37,360 feature vectors, where each vector has 42 attributes.

3. Dimensionality Reduction

SInce the dimensions of the feature vector is large (42 attributes), it was decided to apply dimensionality reduction using the principal component analysis approach (PCA). As a preliminary step, the data was standardized to ensure that all the variables are on the same scale, which is crucial in PCA because it is a variance-based method. If the variables are not on the same scale, some variables with large variances may dominate the analysis, even if they are not the most important for explaining the variance in the data. The PCA was then applied and using the cumulative variance, which is shown in table 1, it was decided that 11 principal components are sufficient since they account for 99% of the variance in the data. The data obtained after PCA will be used in the modeling phase.

Principle Component Number	1	2	3	4	5	6	7	8	9	10	11
Cumulative Variance	0.28	0.54	0.72	0.85	0.94	0.97	0.98	0.99	0.99	0.99	0.99

Table 1. PCA Cumulative Variance

4. Modeling Techniques

Three different types of models were used in order to be able to identify the best model that can classify the signals into their corresponding stages, which are multi-layer perceptron, k-nearest neighbor (KNN), and the random forest classifier. The purpose of using three different techniques is to identify the model that can best classify the signals, where the metric used to compare among the models is the accuracy score of the predicted values.

4.1. Multi-Layer Perceptron

A multi-layer perceptron (MLP) consists of several layers where the input is fed into the network in a sequential manner. For the four stages classification problem, the multi-layer perceptron architecture chosen to develop the model is illustrated in Figure 4. The input layer takes a vector x of dimension 11, the first hidden layer contains 40 neurons, the second hidden layer contains 20 neurons, and the output layer is the classification. The “ReLu” was utilized as the activation function while “adam” was used as the optimizer.

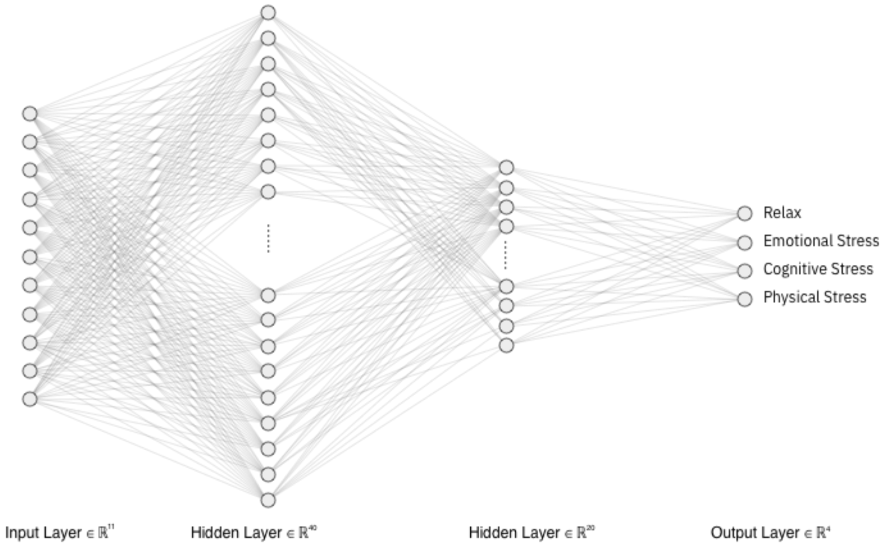
Relax
Emotional Stress
Cognitive Stress
Physical Stress
Input Layer $\in \mathbb{R}^{11}$
Hidden Layer $\in \mathbb{R}^{40}$
Hidden Layer $\in \mathbb{R}^{20}$
Output Layer $\in \mathbb{R}^{4}$

Figure 4. Network Architecture

4.2. KNN

K-nearest neighbors algorithm (KNN) is an instance based supervised learning method that is widely used in classification. Since the data of the project, which are the signals, have labels, which are the four stages, KNN was considered as one of the developed models for the non-EEG signals classification. The KNN algorithm considers the k-nearest neighbors when classifying a point. The model developed in the code considered the nearest 50 neighbors. The model is then fit to the trained set of data, where the training and testing sets will be thoroughly discussed in the validation section, and the testing set is used to test the model and calculate its accuracy.

4.3. Random Forest

The Random Forest classification algorithm is a machine learning ensemble algorithm that uses a lot of smaller decision trees, which are the estimators of the Random Forest classifier, to produce their predictions and then combine all the predictions into a single, more accurate prediction. Random Forest are robust to overfitting and can generalize better than the normal decision trees, which also helps in handling datasets with higher dimensionality. In our model we chose the n_estimators parameter, which is the number of decision trees used in the forest, to be 100, which is the default value of the Random Forest classifier. This choice of parameter gave us the highest accuracy when we tested our trained model on our test data as we did with our previous classifiers. An example of the Random Forest classifier is shown in the image below with multiple trees receiving an instance, predicting the correct class, and finally combining the predictions to get a more accurate final prediction.

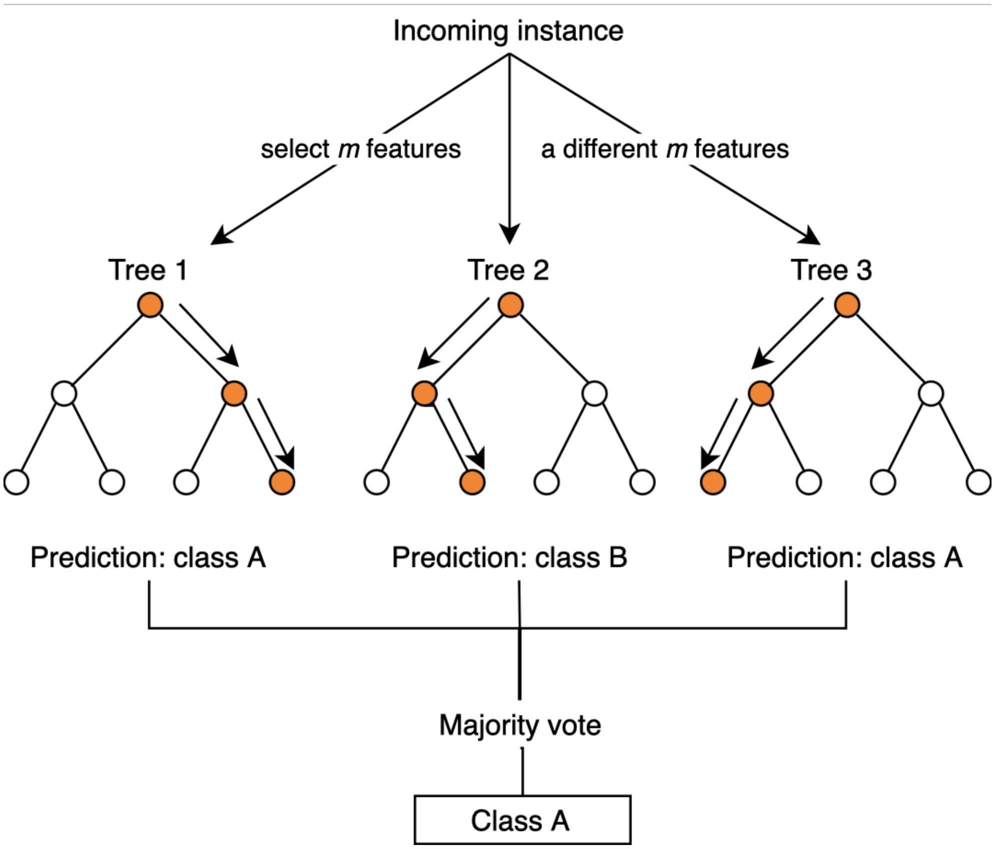

Figure 5. Random Forest Decision Example (from source [2])

5. Validation Techniques

For training and testing a model that can classify the signal as a signal of relaxation, cognitive stress, physical stress, and emotional stress phase, two types of validation were used for each of the previously mentioned models, which are cross-validation per subject and leave one subject-out cross-validation.

5.1. Cross-Validation Per Subject

The cross-validation per subject technique considers only one subject at a time. The data/signals of that specific subject are divided into a training set and testing set. In the developed code, 80% of the subject's signals were used for training the model while the remaining 20% were used for testing the model. The accuracy of the model is then calculated. The previous model was developed for each of the 20 subjects and for each modeling technique. For each of the MLP, KNN, and random forest, the mean accuracy for the 20 subjects were calculated. The results are shown in Table 2.

Modeling Type	Multiple-Layer Perceptron	KNN	Random Forest
Mean Accuracy (%)	98.8	93.7	98.6

Table 2. Cross Validation Per Subject Mean Accuracy

5.2. Leave-one Subject-Out Validation

Leave-one subject-out technique takes into consideration all of the twenty subjects. As inferred by the name of the technique, 19 out of the 20 subjects are used to train the model while the remaining subject is used for testing and the accuracy of the model is then calculated. The same process is repeated for all of the 20 subjects. For instance, in the first iteration, subject one is considered as the test set while the remaining 19 subjects are used to train the mode. In the next iteration, the second subject will be used in the testing phase while the rest are used in training. Similar to the previous validation technique, the mean accuracy for the 20 subjects was calculated for the MLP, KNN, and random forest. The results are shown in Table 3.

Modeling Type	Multiple-Layer Perceptron	KNN	Random Forest
Mean Accuracy (%)	64.5	71.6	77

Table 3. Leave One Subject-Out Validation Mean Accuracy

Conclusion

From tables 2 and 3, it can be concluded that developing models for individual subjects provide higher accuracy scores than developing models that consider all of the subjects at once. In the cross-validation per subject, multi-layer perceptron (MLP) and random forest showed high and similar accuracies. Generally, all modeling techniques showed an accuracy of more than 90% when considering the cross-validation per subject. However, in the leave-one subject-out validation technique, KNN showed higher accuracy than MLP. It can be concluded that when developing customized models, where cross-validation per subject is usually utilized, it is better to develop the model using MLP or random forest. When generalizing the model to increase its scale, KNN is a better approach than MLP along with the random forest method.

References

[1] Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. E215–e220.

[2] Random forests. DeepAI. (2020, September 10). Retrieved December 11, 2022, from https://deepai.org/machine-learning-glossary-and-terms/random-forest