ANALYSIS OF CHILDREN WITH OBSESSIVE-COMPULSIVE-DISORDER USING MACHINE LEARNING

Laurits Fromberg s174512

DTU Compute

`https://github.com/LauritsFromberg/CDA`

1. ABSTRACT

Emotion recognition using wearable devices, as for instance the Empatica E4, has in recent years recieved an increase in popularity within the field of data-analysis research [1]. In particular for predicting the emotional state of individuals based on received biosignal. This can for example be used to assist people who suffer from mental health disorders [2, 3].

This paper examines the use of machine learning models for emotion recognition using biosignals from the Empatica E4, while also analysing the importance of the acquired features. Here uncovering that; it is indeed viable to develop a suitable machine learning model, with a random forest yielding the best performance within the models considered. Moreover, it is in addition distinguished that the features related to the `HR` biosignal are of particular importance.

However, it is also established that a comprehensive search for more appropriate features is deemed necessary, alongside a potential pre-processing of the `BVP` biosignal. Furthermore, an ordinal regression approach may be considered in order to further improve the performance.

2. INTRODUCTION

Obsessive-Compulsive-Disorder is a mental health disorder, which is often associated with excessive hand-washing and an obsession with symmetry, despite being far more extensive than this. It is a mental health disorder which equally affects people across age, sex and ethnicity, hence also children [4]. The most common treatment is (excluding medicine) therapy utilising exposure and response prevention, although this method of treatment ought to be extraordinary difficult for a child to practice outside of therapy thus the need for parental interaction and guidance [4]. This can be assisted by the use of wearable devices such as the *Empatica E4* for classifying when a child will experience obsessions in real-time, to enable the informing of parents to assist in response prevention as established in therapy rather than alleviate compulsive behaviour, which may otherwise be the case. However, prior to such classification being feasible is the requirement for preliminary studies. One preliminary study comprise of assessing the extent, to which it is possible to produce emotion recognition by utilising biosignals from the Empatica E4 [3]. The research question considered are the following:

- How can we develop a machine learning model for emotion recognition using the Empatica E4 biosignals?
- Which biosignals features are here of importance?

From these research question it becomes reasonable to consider different regression models from supervised learning. The models; Elastic Net, Decision Tree, Random Forest and Gradient Boosting are in particular of interest due to their properties of feature importance [5]. Moreover, it ought to be sensible to apply methods to improve performance such as for instance *cross-validation* for hyper-parameter tuning as well as to ensure a proper evaluation of the generalisable properties of the particular models in question [5].

3. DATA

The data considered in this paper is the `WESAD` data-set [6]. This is composed of biosignal measurements from the wearable device Empatica E4, along with labels of emotions with supplementary self-report questionnaires. The study-protocol uses data from 17 participants, two of which were omitted due to sensor malfunctions and one due to containing *Not a Number* (`NaN`) entries. Details now follow in the proceeding.

3.1. Empatica E4

The Empatica E4 is a wrist-worn wearable device that enables real-time acquisition of biosignals thus subsequently allowing emotion recognition to be accomplished directly. It has four sensors recording the following biosignal measurements [2].

PPG Sensor measures the *blood volume pulse* (`BVP`), which is used to derive the *heart rate* (`HR`) in beats per minute (`BPM`) with a sampling-rate of 1 Hz, interbeat interval (`IBI`) and *heart rate variability* (`HRV`).

GSR Sensor measures the *electrodermal activity* (`EDA`), which is variations in electrical properties of the skin.

3-Axis Accelerometer records motion-based activity.

Infrared Thermopile reads the peripheral skin temperature.

Notice here that the abbreviations PPG and GSR stand for; *photoplethysmogram* and *galvanic skin response* respectively.

Table 1: Description of the sensors.

	PPG	EDA	Accelerometer	Thermopile
Rate	64 Hz	4 Hz	32 Hz	4 Hz
Unit	N/A	μS	$1/64g$	$^{\circ}C$

3.2. Experimental Setup

The WESAD data-set is an open-source multimodal data-set for emotion recognition, in particular stress. Here the Empatica E4 constitutes a vital role due to its non-intrusive properties, hence the reason for it often being the preferred measurement device [6]. For a reliable stress recognition system the experimental setup needs to reflect and simulate real-life exposure to stimuli which is stress inducing id est creates a physiological response by the sympathetic nervous system causing the release of hormones, leading to an increase in exempli gratia heart rate and muscle tension [6]. Furthermore, emotions are perceived differently on an individual level, hence why self-report questionnaires are included to represent the subjective experience of each individual and to enable the development of potentially personalised models [6]. The data collection procedure is implemented by the participants wearing the Empatica E4 on their non-dominant wrist and then elicit three different emotional states within the participants; neutral (baseline), stress and amusement, with the addition of a guided meditation hereafter to calm the participants. The study-protocol now follows [6]

Preparation: The subject are advised to avoid caffeine and demanding exercise. The participants are equipped with the Empatica E4 and conduct a short test.

Baseline: A 20 minute baseline is recorded with the participants sitting or standing with access to neutral reading material in order to elicit a neutral state.

Amusement: The subject watch a total of 11 amusing video clips with each clip being followed by a short neutral sequence of 5 seconds. In total 392 seconds.

Stress: The subjects are exposed to the well-established *Trier Social Stress Test* (TSST) (refer to [7] for details) consisting of public speaking and mental arithmetic. The public speaking involved a five minute speech on their personal traits (strengths/weaknesses) in front of a three-person panel. The subjects have three minutes of preparation. Proceeding with mental arithmetic, with the subjects counting from 2023 to zero using a stride of 17, restarting for each mistake. In total 10 minutes with an equal amount of rest hereafter.

Meditation: The subjects are guided through a controlled breathing exercise of seven minutes by an audio track. This is done after both stress and amusement.

Recovery: The equipment is removed. The two main emotional states stress and amusement are interchanged between participants to ensure that the order has no effect.

3.3. Self-Report Questionnaires

Succeeding each of the aforementioned emotional states the participants answer a self-report questionnaire in order to validate that the desired emotional state is indeed elicited as well as to enable the development of personalised models. Hereby obtaining a total of five self-report questionnaires for each participant, consisting of selected well-established questionnaires. The first being the *Positive and Negative Affect Schedule* (PANAS) questionnaire, which contains 20 questions; 10 for positive affects and 10 for negative affects respectively, each rated on a five point *Likert scale* [8]. Four additional questions are added; *stressed?*, *frustrated?*, *happy?* and *sad?*. Furthermore, adding six questions from the *State-Trait Anxiety Inventory* (STAI) to acquire insight into the anxiety level of the participants, which here uses a four point Likert scale (please refer to [6, 9]). A *Self-Assessment Manikins* (SAM) questionnaire is likewise included, in order to deduce the level of *valence* and *arousal* using a nine point scale (1=low and 9=high) [10]. Finally, following the emotional state of stress, nine questions from the *Short Stress State Questionnaire* (SSSQ) are utilised to identify the type of stress using a five point Likert scale, although three of these questions are included in the PANAS questionnaire with two only being asked after the emotional state of stress [11].

For more details in relation to the questions please consult [6].

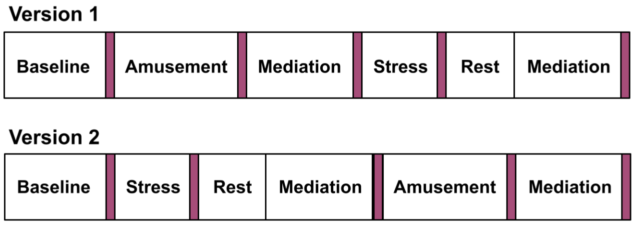

Fig. 2: Overview of the two versions of the study-protocol.

The objective is now to predict the answers to the self-report questionnaires given the biosignals from the phase (emotional state) prior, hence yielding a supervised learning problem. This is done for all self-report questionnaires except the short stress state questionnaire due to this only being applicable to the phase of stress. However, due to the serial correlation of the data, we perform feature extraction by utilising

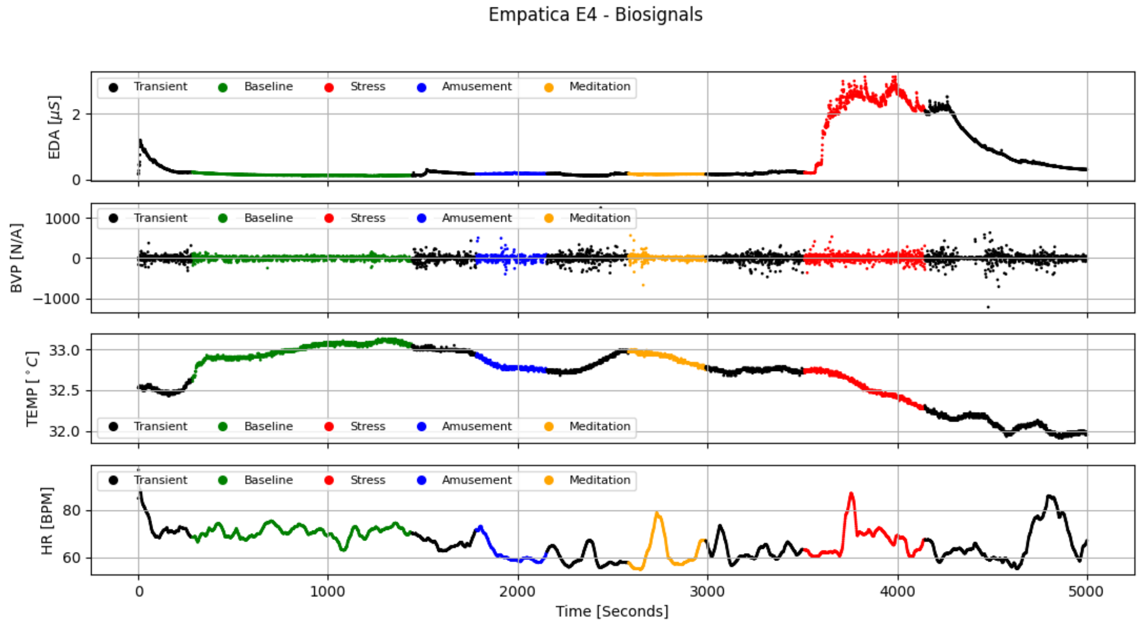

Fig. 1: Example of biosignals; for subject number four.

the *sliding-window* method in order to use classical machine learning models and procedures, which require the need for independent and identically distributed data [9].

Note, that the labels of emotions are used to divide the data into the respective phases and since these have a sampling-rate of 700 Hz, we ought to re-sample BVP to 70 Hz. This being due to the sampling-rates otherwise not being divisible, hence it becomes infeasible to divide BVP into the respective phases with reference to the labels. Furthermore, note that the signals from the 3-axis accelerometer are discarded due to not being of interest for this particular study.

3.4. Pre-Processing

The pre-processing exploits a five step procedure, which here is composed of the following five steps [6, 12, 13, 14, 15]:

Filtering: The EDA biosignal is filtered with a low-pass filter such as the *Butterworth* filter, as proposed by [12]. Here using a 4'th order Butterworth filter with a cut-off frequency of 2.5 Hz in order to reduce motion artifacts.

Smoothing: The EDA biosignal is hereafter smoothed with an exponential moving-average as proposed by [13]. The smoothing is performed using a smoothing factor of $\alpha = 0.4$, to further reduce motion artifacts and noise.

Decomposition: The EDA biosignal is then decomposed using the cvxEDA algorithm as proposed by [14] and implemented in [15]. This algorithm decomposes the EDA biosignal into a phasic and a tonic component with the phasic component reflecting short-time response to stimuli, while the tonic component expresses slow drift of the baseline as well as spontaneous change [14].

Segmentation: Each biosignal is then separated into multiple segments, by utilising the sliding-window method. This is done by letting a window slide through each signal, using a window-size of 65 seconds and a window-shift of 13 second, resulting in overlapping segments [6]. Further applying a burn-in period of 10 seconds.

Feature Extraction: Finally, we extract a set of features for each of the respective segments, which is done for all signals. The pre-specified procedure of applying the minimum, the maximum and the sample-mean is then employed in order to produce a set of overall features for the entire signal, which is done for all signals [6]. Please refer to table 5 for an overview of the features.

4. METHODS

A number of classical supervised machine learning models for regression have been applied to the considered problem; *Elastic Net* (EN), *k-nearest neighbours* (KNN), *Decision Trees* (DT), *Random Forests* (RF) and *Gradient Boosting* (GB).

Here the Elastic Net, the Decision Trees, the Random Forests as well as the Gradient Boosting are all chosen due to their feature importance properties in order to assess which biosignal features are of importance. The last two being ensemble models, which are more complex and sometimes yield better performance. The k-nearest neighbours is chosen due to being simple and thus constituting a good and robust baseline for model comparison and selection purposes. We shall now subsequently present the different models in further details.

4.1. Models

A brief introduction to the models will now be given. Elastic Net is a linear regression model, which is regularised by a linear-combination of the $\mathcal{L}_1$ and $\mathcal{L}_2$- norm, specified by

$$\inf_{\beta \in \mathbb{R}^p} \frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \left(\frac{1}{2}(1-\alpha)\|\boldsymbol{\beta}\|_2^2 + \alpha \|\boldsymbol{\beta}\|_1 \right). \quad (1)$$

The feature importance properties stems from Elastic Net being able to set coefficients to exactly zero. Furthermore, it has the useful property of being able to fit models where $p > n$, which here is the case. Notice that both *Ridge regression* as well as *Lasso* are special cases hereof [5].

k-nearest neighbours is a non-parametric model, which makes predictions by taking the sample-mean of the associated k-nearest neighbours, which is resolved with reference to the Euclidean distances. This is implemented with a weighting proportional to the Euclidean distance [5].

Decision trees often referred to as CART (*classification and regression trees*) is a class of simple non-parametric models where the feature space is subdivided into a number of partitionings and then fits a constant in each interval and can thus be viewed as a piecewise constant approximation. The splitting is done by minimising for instance the *sum-of-squares*, hence the optimal constant function being the sample-mean. The splitting is done until an interval only contains a prespecified number of observations. for more details confer [5].

Random forests constructs multiple decision trees and then in the regression case, makes a prediction based on the sample-mean of all the predictions from the decision trees [5].

Gradient boosting is an ensemble method, typically consisting of decision trees similar to random forests. The model is build in a forward stage-wise additive fashion, where a decision tree is fitted to the negative gradient and any differentiable loss-function is permitable (for more details see [5]).

4.2. Cross-Validation

In order to estimate the generalisation errors of each model a *leave-one-patient-out cross-validation* scheme is employed due to the amount of data available. Here using 10 splits, where in each split a single subject chosen at random is left out for testing and the models and the training of the models is executed using the remaining subjects. Each split then applies a gird-search using 5-fold cross-validation for hyper-parameter tuning. Here using a rather coarse grid due to the time complexity associated, implicating severe limitations. This is done to determine the generalisable performance of each model by taking the sample-mean over all test-errors, which is done for model selection purposes. Notice that the input features have been standardised prior to the training of the respective models as proposed by [5].

4.3. Performance Measures

For evaluating the performance of each model in the cross-validation procedure, we consider the *mean-squared-error* (MSE), while for the assessment of the final model configured through the use of the cross-validation, we additionally examine the *mean-absolute-error* (MAE) as well as the $\mathcal{L}_\infty$-norm in order to provide a variety of measures for greater insight.

$$\mathtt{MSE} = \frac{1}{n} \sum_{i=1}^{n} \|\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i\|_2^2 . \quad (2)$$

$$\mathtt{MAE} = \frac{1}{n} \sum_{i=1}^{n} \|\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i\|_1 . \quad (3)$$

$$\mathcal{L}_\infty = \sup_{i,j} |y_{ij} - \hat{y}_{ij}|. \quad (4)$$

4.4. Feature Importance

An assessment of the feature importance for models using decision trees can here be established by computing the normalised (total) reduction in the specified splitting criterion for the feature under consideration, which is known as the *Gini importance* [5, 6]. For random forests this is done with a similar approach but where the sample-mean is applied across the trees. Feature importance can likewise be done for boosting. Note that we restrict the investigation to comprise of the features that have a Gini importance, which are greater in magnitude than the *standard-error of the mean* id est $\sigma/\sqrt{n}$.

5. RESULTS

The cross-validation procedure resulted in the following generalisation errors, for each of the respective models.

Table 2: Generalisation Errors.

Model	Gen. Error
EN	28.391
KNN	29.079
DT	35.283
RF	$\boxed{25.877}$
GB	28.843

It is hereby seen that the random forests produces the smallest generalisation error, hence this is chosen as the final model. The best hyper-parameters found in the grid-search using

5-fold cross-validation were the following (corresponding to `sklearn.ensemble.RandomForestRegressor()`).

Table 3: Hyper-parameters.

max_depth	3
max_features	50
min_samples_leaf	1
min_samples_split	4
n_estimators	90
random_state	10220

Here the random state is provided for reproducability reasons.

Preceding the feature importance analysis, a final assessment of the random forests model is completed by training the model on all the training data as realised in [16] and then evaluating it using the described performance measures.

Table 4: Final Assessment

MSE	MAE	$\mathcal{L}_\infty$
17.053	17.004	2.435

The final assessment appears to be appropriate, with the model achieving adequate performance compared to picking a score at random, for each respective scale [6].

Further analysis consists of the feature importance in table 5.

6. DISCUSSION

From the proceeding section it is evident that the best model is the random forests, which may not be surprising as the targets are represented by the integer values encountered through the individual scales, while a decision tree constitutes a piecewise constant approximation, hence an ensemble of such models ought to be a sensible model choice for said data [5]. Despite this, ordinal regression may provide additional benefits to the otherwise standard regression approach. However, further work needs to be accomplished in order to establish whether or not this is in fact the case.

From the feature importance it is observed that a majority of the features are not particularly important, especially those associated with `BVP` and `TEMP`. This may be due to these biosignals (as seen in figure 1) either containing a lot of noise or being close to constant. Here `TEMP` is close to constant, hence not providing a lot of information, whereas `BVP` to a greater extent has a lot of noise (low-signal-to-noise-ratio), alongside extreme outliers (which is the reason for the axis). Please consult figure 3 to observe the appearance of noise.

Additional pre-processing may therefore yield better results. It is also noticeable that a lot of the features in general do not provide much gain, which perhaps ought to be explained by the following arguments;

- This study uses rather elementary features, which may not be suitable for the specific biosignals and as such more domain specific features ought to be considered.
- The same features are considered for all biosignals, albeit this not necessarily being the preferred approach due to specific biosignals perhaps necessitating specific and thus likely also more complex features.

A search for more optimal as well as operative features should therefore be at its place and considered for future work. This could for instance include, although is not limited to, domain specific features [17], frequency specific features, time-series features; auto-correlation, Fourier features, et cetera.

Lastly, the features that are in fact of importance includes for instance the `HR` features such as the minimum standard deviation, the minimum entropy and the minimum absolute integral. The `HR` biosignal thus appears to be rather fundamental as one may suspect from a physiological standpoint (the heart-rate differs between calm and stressful situations). Furthermore, the phasic component of the `EDA` biosignal also seems to be more vital than its tonic counterpart, which may likewise be demonstrated on account of physiological grounds; as the phasic part reflects short-time response to stimuli rather than a drift of the baseline as it may be the case during stressful circumstances.

Note that the features of importance are primarily associated to the minimum values and thus perhaps indicating that the model utilises this information to predict non-stressful states as these are predominant and conceivably related to these.

7. CONCLUSION

In conclusion; it is indeed possible to develop a machine learning model for emotion recognition using the biosignals of the Empatica E4, with the random forests here being the most suitable model within those considered. Furthermore, it was likewise feasible to analyse the biosignal features that were of importance, which deemed to primarily be those associated with the `HR` biosignal.

However, supplementary inspections using for instance a more extensive grid-search, more data and perhaps ordinal regression methods ought to improve the results. Moreover, a comprehensive study of appropriate and suitable features would perhaps yield a significant breakthrough, alongside a pre-processing of the `BVP` biosignal.

8. REFERENCES

[1] S. Saganowski, “Bringing emotion recognition out of the lab into real life: Recent advances in sensors and

Table 5: Feature Importance.

Features	EDA		TEMP	BVP	HR
	(tonic)	(phasic)			
max. slope	0.0018	N/A	0.0024	N/A	N/A
min. slope	0.0030	N/A	0.0033	0.0029	0.0017
sample-mean slope	N/A	0.0070	N/A	0.0086	0.0020
max. avg.	0.0022	0.0047	N/A	0.0026	0.0029
min. avg.	N/A	0.044	N/A	N/A	0.0019
sample-mean avg.	0.0040	0.0131	N/A	0.0077	0.0019
max. median	0.0037	0.0072	N/A	0.0250	0.0029
min. median	0.0027	0.0049	N/A	N/A	N/A
sample-mean median	N/A	0.0292	0.0025	0.0091	N/A
max. std.	0.0033	0.0029	0.0062	N/A	0.0058
min. std.	0.0021	N/A	N/A	0.0057	0.0771
sample-mean std.	0.0109	0.0020	N/A	N/A	0.0083
sample-mean min.	0.0038	0.0590	0.0017	0.0089	0.0029
sample-mean max	0.0032	0.0090	0.0055	0.0195	0.0022
min. entropy	0.0045	0.0018	0.0023	N/A	0.1345
max. entropy	0.0051	0.0166	0.0036	N/A	N/A
sample-mean entropy	0.0062	N/A	0.0093	N/A	0.0139
max. avg. grad.	0.0271	0.0143	0.0059	0.0026	0.0026
min. avg. grad.	0.0047	0.0198	0.0071	0.0021	0.0023
sample-mean avg. grad.	0.0387	0.0267	0.0148	N/A	0.0095
max abs. integral	0.0033	0.0041	N/A	0.0035	N/A
min abs. integral	0.0038	0.0107	0.0068	N/A	0.0713
sample-mean abs. integral	0.0035	0.0282	N/A	0.0470	0.0052

machine learning," *MDPI*, 2022.

[2] Empatica e4. [Online]. Available: https://www.empatica.com/en-int/research/e4/

[3] F. Sabry, T. Eltaras, W. Labda, K. Alzoubi, and Q. Malluhi, "Machine learning for healthcare wearable devices: The big picture," *Hindawi*, 2022.

[4] International ocd foundation. [Online]. Available: https://iocdf.org/

[5] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Predictions*, ser. Texts in Statistical Science. Springer, 2017.

[6] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. V. Laerhoven, "Introducing wesad, a multimodal dataset for wearable stress and affect detection," *Association for Computing Machinery*, 2018.

[7] C. Kirschbaum, K.-M. Pirke, and D. H. HellHammer, "The 'trier social stress test' - a tool for investigating psychobiological stress responses in a laboratory setting," *Neuropsychobiology*, 1993.

[8] D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: The panas scales." *Journal of Personality and Social Psychology*, 1988.

[9] B. H. Barker, H. R. B. Jr., and A. P. W. Jr., "Factor analysis of the items of the state-trait anxiety inventory," *Journal of Clinical Psychology*, 1977.

[10] J. D. Morris, "Observations: Sam: The self-assessment manikin an efficient cross-cultural measurement of emotional response 1," *Journal of Advertising Research*, 1995.

[11] W. S. Helton and K. Naswall, "Short stress state ques-

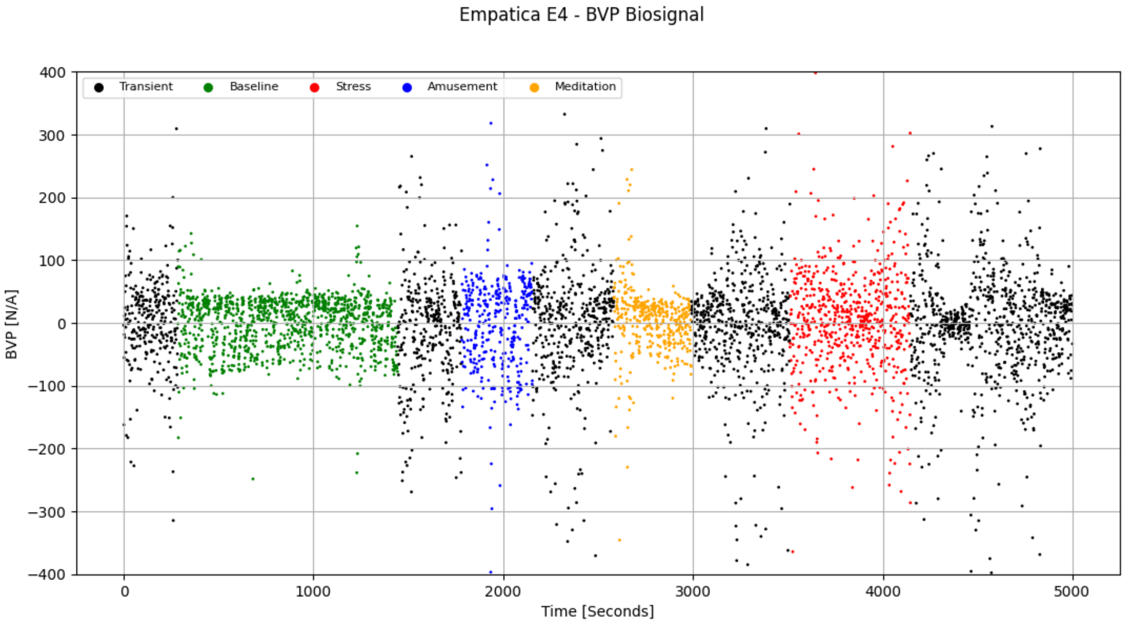

Fig. 3: Example of the `BVP` biosignal. Here excluding extreme outliers.

tionnaire," *European Journal of Psychological Assessment*, 2014.

[12] M.-Z. Poh, T. Loddenkemper, N. C. Swenson, S. Goyal, J. R. Madsen, and d Rosalind W. Picard, "Continuous monitoring of electrodermal activity during epileptic seizures using a wearable sensor," *IEEE*, 2010.

[13] J. Hernandez, R. R. Morris, and R. W. Picard, "Call center stress recognition with person-specific models," *ACII*, 2011.

[14] A. Greco, G. Valenza, A. Lanata, E. P. Scilingo, and L. Citi, "cvxeda: a convex optimization approach to electrodermal activity processing," *IEEE transactions on bio-medical engineering*, 2016.

[15] D. Makowski, T. Pham, Z. J. Lau, J. C. Brammer, F. Lespinasse, H. Pham, C. Scholzel, and S. H. A. Chen, "Neurokit2: A python toolbox for neurophysiological signal processing," *Behavior Research Methods*, 2021.

[16] L. K. H. Clemmensen, S. Das, and N. L. Lund. Analysis of emotions using physiological signals: a pilot study. [Online]. Available: https://bit.ly/3SiJkSK

[17] J. Shukla, M. Barreda-Ángeles, J. Oliver, G. C. Nandi, and D. Puig, "Feature extraction and selection for emotion recognition from electrodermal activity," *IEEE Transactions on Affective Computing*, 2019.