Towards a Minimalistic Stress Classification Method based on HRV

Roswitha Duwenbeck
Systeme der Medizintechnik
Universität Duisburg-Essen
Duisburg, Germany
E-mail: roswitha.kressner@uni-due.de

Elsa Andrea Kirchner
Systeme der Medizintechnik & Robotics Innovation Center
Universität Duisburg-Essen & German Research Center for Artificial Intelligence
Duisburg, Germany & Bremen, Germany
E-mail: elsa.kirchner@uni-due.de & elsa.kirchner@dfki.de

***Abstract*—Stress is a feeling of emotional and physical tension, that poses as a risk factor in many diseases, for example the nervous, musculoskeletal, cardiovascular or gastrointestinal system. Fast and easy detection could be a first step in order to help people manage their stress-levels. This paper depicts an ongoing work in the domain of stress prediction with Heart Rate Variability related features by classifying two different levels on the Stress-Predict Dataset. The performance of different classifiers was tested with Leave-One-Subject-Out Cross Validation and compared to each other. The best performance was reached with the Aggregated Mondrian Forest Classifier and a mean balanced accuracy of 97.87%.**

***Index Terms*—*Heart rate variability*; *stress prediction*; *Machine Learning*.**

I. Introduction

There are manifold definitions of stress in humans. Maybe the most generic and well-known definition is the one by Hans Seyle, stating that ”Stress is the non-specific response of the body to any demand” [1]. This demands can be of physical or psychological nature [2] [3]. Especially psychological stress, which can be defined as ”stress that occurs when an individual perceives that environmental demands tax or exceed his or her adaptive capacity” [4], is a topic of great interest in healthcare. While both stressors, psychological and physiological, are of very different nature and by thus, have different effects on the body, they also share a subset of comparable effects on the body [5]. Not only do they cause similar subjective and hormonal responses, it has also been suggested that they also share common neural substrates [6]. Long term effects of both stressors seem to show more differing symptoms on the body. While effects of too much psychological stress are well documented and mostly about harm of the nervous, musculoskeletal, respiratory, cardiovascular, gastrointestinal, reproductive, and other systems [7], the influences of physical stress are not so clearly outlined. Nevertheless there are studies, which observed effects like increased luminal permeability [8] or differences in corticosterone serum-levels [9]. Short term effects of both stressors tend to be more similar, but still differing. Both stress types cause a rise in physiological parameters like Heart Rate (HR), Breathing Rate or Oxygen Consumption, but differ in severity. Like this, the Oxygen Consumption and Breathing Rate is higher in physical than in psychological stress. The HR, on the other hand, is higher in psychological stress. [11]

The possible damages outline the necessity of a reliable stress detection method. Can et al. state, that researchers found out that stress should be handled when the symptoms first come out to avoid the long-term consequences [12]. This can be important in different settings like the workplace, traffic and generally in healthcare.

Because of its importance, automated stress detection is not a new topic in the area of Machine Learning (ML). There are already many different approaches in terms of using psychophysiological signals, selected features, and machine or deep learning methods. Can et al. used different ML methods, such as Linear-Discriminant Analysis (LDA), Support-Vector-Machine (SVM), k-nearest-Neighbors (kNN), Logistic Regression (LR), Random Forest (RF) and Multi Layer Perceptron (MLP) to detect 3 psychological stress-classes with the help of HR, Electrodermal Activity (EDA), Inter Beat Intervalls (IBI), Skin Temperature (ST) and Acceleration [12]. Costin et al. used Heart Rate Vsariability (HRV) related features from the Electrocardiogram (ECG), to train a Minimum Distance Classifier (MDC) and detect three psychological levels of stress [13]. Garg et al. used ECG, body temperature (TEMP), Respiration (RESP), Electromyogram (EMG), and EDA to classify two or three physiological conditions - neutral (baseline), psychological stress and neutral (baseline), psychological stress, amusement with the help of kNN, LDA, RF, AdaBoost (AB), and SVM [14]. Their best

TABLE I
Selection of papers detecting Stress with ML.

Paper	**Details**		
	Signals	***ML-Method***	***Best Results***
[12]	HR, EDA, IBI, ST, Acceleration	LDA, SVM, kNN, LR, RF, MLP	Accuracy of 97.92%
[13]	ECG (HRV)	Minimum Distance Classifier	Accuracy of 89.36%
[14]	ECG, TEMP, RESP EMG, EDA	kNN, LDA, RF AB, SVM	Accuracy of 84.17%
[15]	ECG (HRV)	kNN, SVM, MLP RF, GB	F1 of 79%
[16]	ECG(HRV)	SVM, MLP, IBK DT, LDA	Accuracy of 94%

 ISBN: 978-1-68558-100-8

classification results were reached in the binary classification task with RF and 84.17% Accuracy, while in the three class problem they reached 67.56% Accuracy, also with RF [14]. Dalmeida et al. classified psychological stress in two classes with ECG-derived HRV-Features and different ML methods such as kNN, SVM, MLP, RF and Gradient Boosting (GB) [15]. Castaldo et al. classified two classes of psychological stress by using ECG derived HRV-Features with SVM, MLP, Neighbor Search (IBK), DT and LDA [16].

It becomes clear that the classification of stress can be done with the help of many psychophysiological features. And although using HRV traits alone does not appear to be as accurate as combining multiple signals, it is nevertheless an interesting approach for minimal applications. This paper can be seen as a starting point in creating a minimalist stress classification method, which is robust and practical for the in use real world scenarios. This paper is structured as follows: In Section II the used materials, the Dataset and it´s preparation, the chosen features and the ML methods and validation procedure are explained. Section III describes the validation results, while Section IV draws conclusions out of the results and lists future plans for the project, as this paper depicts an ongoing work.

II. Methods and Materials

The Stress-Predict Dataset (SPD) was used to train different ML classifiers in a binary classification task, classifying a stressed and a rest state. To make training possible, steps such as preparing the dataset, feature extraction and choosing the ML methods had to be done. These steps, and also the used materials, are described in here.

A. The Dataset

The relatively new SPD from Iqbal et al. consists of bio-signals from 35 participants, 25 men and 10 women. Stressors were forced Hyperventilation, the Trier Social Stress Test and Stroop Color Test. An E4 watch from Empatica was used to measure individual physiological changes based on PPG. The signal was filtered to get a clean Blood Volume Pulse (BVP), which was used to obtain the HR, IBI and Respiration Rate (RESP) by an estimation algorithm. [18]
To the authors knowledge this dataset was not used before to detect stress from HRV features.

B. Preparation of the dataset and preprocessing

Because the dataset was not mainly composed to detect stress from HRV-Features, but from HR and RESP, an own assignment of labels, timestamps and physiological parameters had to be done. Iqbal et al. original distribute one "processed" data folder and one with raw data. The first folder contains a list, with merged patient label, HR, RESP a stress-label and a timestamp for every second, given in ms with one decimal place. The raw data contains separate lists of the physiological signals with the passed time since the start, given in ms with 6 decimal places, and the starting time of the experiment in ms as header, for every subject.

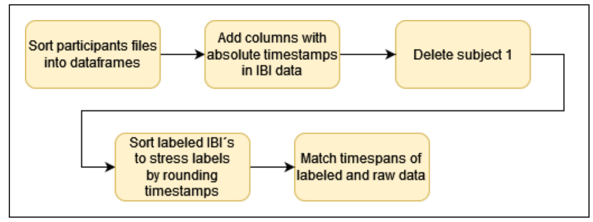

Fig. 1. Dataset preparation

It was important to assign labels to the IBIs. At the start, subject data was sorted into dataframes. It could be seen, that subject one has no matching labeled data, so it was deleted. The raw condition of the IBI dataframes just contains the IBI´s and the passed time in ms from the respective starting time. To facilitate sorting labels to IBI´s, a new column for the absolute passed time since start, as given in the processed filed, was added. The starting time was given in the raw header file, by this the columns could be filled with an iterative addition of the start time and the passed time. Also, the processed data is time-wise longer than the raw IBI data. Processed data which could not be associated to any IBI data was therefore deleted. From there on, the labeled data was sorted to each IBI by rounding the IBI-times: Each raw timestamp, given in ms, was rounded to match a processed timestamp given in seconds, and by this sorted to one Stress-Label. The process can be seen in Figure 1. In a next step, the IBI-signals were windowed into 60-sec-windows. Since the windowing sometimes produced windows with two different labels, the window was labeled according to the majority of stress-labels. Windows with less than 30 IBI's were sorted out as it would not be physiological.

C. HRV Features

To extract HRV Features from IBI-signals the pyhrv-toolkit was used [17]. The chosen features can be sorted in time- and frequency-domain-, but also nonlinear features.

1) Time Domain Features: Time domain features included "NN" Parameters, which denote the time between two consecutive R-peaks in an ECG signal [19]. Different statistical features like the NN-Counter, mean, minimum and maximum of the time window and differences in the time window were taken. Furthermore the standard deviation of the NN´s was taken, the standard deviation (SD) of the average NN, the root mean square of successive differences, the number of pairs of successive NN´s that differ by more than 50 ms (nn50) and 20ms (nn20), the proportion of NN50 and NN20 divided by total number of NN´s (pNN50 and pNN20).

2) Frequency Domain Features: To obtain frequency domain features, Welch´s Power Spectral Density was used. For the classification task, the absolute powers of the very low (0.00Hz - 0.04Hz), low (0.04Hz - 0.15Hz) and high frequency (0.15Hz - 0.40Hz) band was used. Also, the total power of all frequency bands and the ratio of the power of the low and high frequency bands.

TABLE II
TESTED ML METHODS

Classifier	Parameters
Dummy Classifier	strategy = most frequent
Multi Layer Perceptron	max_iter=45, hidden_layers=45, 20, batch_size=15
Passive Aggressive Classifier	C=0.0, fit_intercept=False, early_stop=True, max_iter=50
SGD Classifier	penalty='l2', alpha=0.01, max_iter=100, eta0=0.1, epsilon=0.01, early_stop=True
Support Vector Machine One Vs. Rest	C=100.0, degree=10
Gaussian Naive Bayes	All standard
Decision Tree	criterion=entropy
Random Forest	All standard
Support Vector Machine One Vs. One	C=100.0, degree=10
Hoeffding Adaptive Tree	grace_period=100, delta=1e-5, seed=0 leaf_prediction='nb', nb_threshold=10
Hoeffding Tree	grace_period=100, delta=1e-5, binary_split=True
Aggregated Mondrian Forest	n_estimators=5, seed=45
Adaptive Random Forest	n_models=7, seed=45

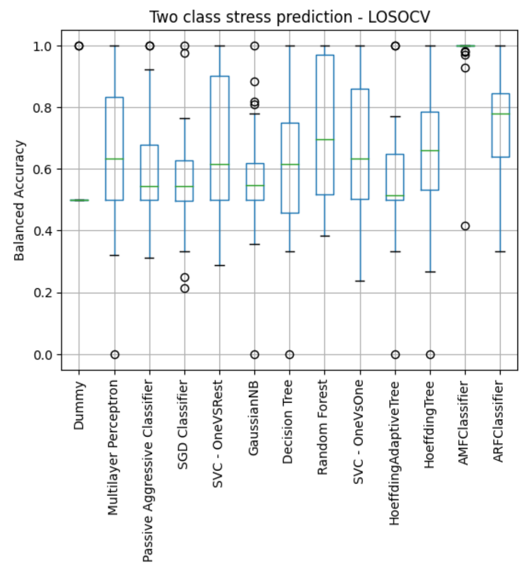

Fig. 2. Boxplots of balanced accuracies

3) Nonlinear Features: Chosen nonlinear features were SD1 and SD2, which are the SD of the data series along the minor axis and the major axis of the Poincaré-Plot.

D. ML-Methods and Learning

A variety of ML methods were used to find the best method for this use case. Used classifiers with their parameters can be found in Table II. The first nine classifiers are from scikit-learn [20], while the last four were taken from river [21]. There was no Hyperparameter Tuning, parameters were rather chosen by experience or trial runs. To gain a better understanding of the robustness of each classifier with respect to completely unseen data, a Leave-One-Subject-Out Cross Validation (LOSOCV) algorithm was written. This own implementation was mainly necessary, because the online learning behaviour of the classifiers based on the river library is hardly compatible with scikit-learn. Before training and testing the data was artificially balanced with Synthetic Minority Oversampling and scaled with a Standard-Scaler.

III. Experimental Results

Boxplots of the mean balanced accuracies are shown in Figure 2. The x-axis shows used classifiers, the y-axis shows the mean balanced accuracies across all subjects. Detailed results can be seen in Table III. Listed are the means over all subjects. It is obvious, that the Aggregated Mondrian Forest (AMF) outperforms the other classifiers by far and keeps up with the state of the art seen in [16].

Lessons learned are that ensemble methods, like RF or the SVM-Methods performed better than non-ensemble learners. Also, non-linear methods, like the Ensemble-SVM´s with radial basis function-kernels, outperformed linear methods, like the Passive Aggressive- or SGD-Classifier. It is generally known that both methods perform better on a large number of samples and/or characteristics, so this is not surprising. The Dummy Classifier performed as expected with one outlier being a subject that only expresses the state of no stress, because of the out-sorting of windows with less than 30 IBI´s.

IV. Conclusion and Future Work

The AMF was able to classify the two states "No Stress" and "Stress" with nearly perfect results. Why exactly this classifier outperforms others by a large margin, has to be investigated. To this moment at least incorrect infusion of the output label in the testing data has been ruled out. Because (A)MFs are relatively new [21], there is not much literature about best use cases or pro and contra. Generally, MFs seem to prefer sparse feature spaces [21], which is the case here. The authors hope to generate an answer about reasons for the excellent performance in the near future.

To receive an answer regarding the unusal good performance of the AMFs, a future task would be to test the same classifiers on a different dataset. Currently, efforts are being made to use "MIT DriveDB" [23] to gain more knowledge about AMFs, while also testing classification accuracy with more than two classes. The aim of more than two stress classes is to split up them in various intermediate stress levels, as it better fits real-world applications.

An additional goal is to investigate more than one time window, because different window lengths seem to result in varying accuracies. At the moment, 5 minute windows still seem to be the recommended ones [24]. In addition, the use of an individual baseline could be useful: Since each person has slightly different characteristics of HRV related to factors such as age, health, etc., it could lead to better comparability. Furthermore a classical Hyperparameter Tuning could lead to better results for all methods. Finally, as in [7], a long-term goal would be to distinguish between the causes for stress, psychological or physical, to gain better insight into the causes and possible effects of stressful events.

ISBN: 978-1-68558-100-8

ACKNOWLEDGMENTS

This work is funded by the Federal Ministry for Economic Affairs and Climate Action and the German Aerospace Center (DLR) [grant number 50RP2260A].

APPENDIX

TABLE III
CLASSIFICATION PERFORMANCES (LOSOCV)

Classifier	Performance
Dummy Classifier	Accuracy: 62.65%
	Balanced Accuracy: 52.94%
	Precision: 34.26%
	Recall: 52.94%
Multi Layer Perceptron	Accuracy: 65.62%
	Balanced Accuracy: 65.63%
	Precision: 63.30%
	Recall: 64.30%
Passive Aggressive Classifier	Accuracy: 59.62%
	Balanced Accuracy: 59.53%
	Precision: 56.02%
	Recall: 57.97%
SGD Classifier	Accuracy: 55.03%
	Balanced Accuracy: 56.20%
	Precision: 53.43%
	Recall: 55.71%
Support Vector Machine One Vs. Rest	Accuracy: 67.00%
	Balanced Accuracy: 66.42%
	Precision: 63.13%
	Recall: 65.08%
Gaussian Naive Bayes	Accuracy: 53.55%
	Balanced Accuracy: 56.58%
	Precision: 50.24%
	Recall: 55.38%
Decision Tree	Accuracy: 63.96%
	Balanced Accuracy: 63.13%
	Precision: 63.15%
	Recall: 62.06%
Random Forest	Accuracy: 71.67%
	Balanced Accuracy: 71.96%
	Precision: 69.68%
	Recall: 71.96%
Support Vector Machine One Vs. One	Accuracy: 67.21%
	Balanced Accuracy: 67.21%
	Precision: 63.54%
	Recall: 65.43%
Hoeffding Adaptive Tree	Accuracy: 53.81%
	Balanced Accuracy: 57.37%
	Precision: 48.34%
	Recall: 56.88%
Hoeffding Tree	Accuracy: 61.27%
	Balanced Accuracy: 65.12%
	Precision: 62.61%
	Recall: 63.30%
Aggregated Mondrian Forest	Accuracy: **97.79%**
	Balanced Accuracy: **97.87%**
	Precision: **97.50%**
	Recall: **97.87%**
Adaptive Random Forest	Accuracy: 69.58%
	Balanced Accuracy: 73.89%
	Precision: 68.12%
	Recall: 72.19%

REFERENCES

[1] H. Seyle, Stress in health and disease, Butterworth-Heinemann, p. 15, 2013.

[2] Y. Li, J. Qin, J. Yan, N. Zhang, Y. Xu, Y. Zhu, L. Sheng, X. Zhu and S. Ju, “Differences of physical vs. psychological stress: evidences from glucocorticoid receptor expression, hippocampal subfields injury, and behavioral abnormalities,” Brain Imaging And Behavior, vol. 13, pp. 1780-1788, 2019.

[3] S. Rao, R. Hatfield, J. Suls and M. Chamberlain, “Psychological and physical stress induce differential effects on human colonic motility,” The American Journal Of Gastroenterology, vol. 93, pp. 985-990, 1998.

[4] S. Cohen, R. Kessler and L. U. Gordon, “Strategies for measuring stress in studies of psychiatric and physical disorders,” in Measuring Stress: A Guide For Health And Social Scientists, vol. 28, pp. 3-26, 1995.

[5] Y. Nakatake, H. Furuie, M. Yamada, H. Kuniishi, M. Ukezono, K. Yoshizawa and M. Yamada, “The effects of emotional stress are not identical to those of physical stress in mouse model of social defeat stress,” Neuroscience Research, vol. 158, pp. 56-63, 2020.

[6] L. Kogler, V. Müller, A. Chang, S. Eickhoff, P. Fox, R. Gur and B. Derntl, “Psychosocial versus physiological stress—Meta-analyses on deactivations and activations of the neural correlates of stress reactions,” Neuroimage, vol. 119, pp. 235-251, 2015.

[7] K. Hong, “Classification of emotional stress and physical stress using a multispectral based deep feature extraction model,” Scientific Reports, vol. 13, p. 2693, 2023, doi: 10.1038/s41598-023-29903-3.

[8] J. Soederholm and M. Perdue, “II. Stress and intestinal barrier function,” American Journal Of Physiology-Gastrointestinal And Liver Physiology, vol. 280, G7-G13, 2001, doi: 10.1152/ajpgi.2001.280.1.G7.

[9] A. Kavushansk, D. Ben-Shachar, G. Richter-Levin and E. Klein, “Physical stress differs from psychosocial stress in the pattern and time-course of behavioral responses, serum corticosterone and expression of plasticity-related genes in the rat,” Stress, vol. 12, pp. 412-425, 2009.

[10] R. Costin, C. Rotariu and A. Pasarica, “Mental stress detection using heart rate variability and morphologic variability of EeG signals,” The 2012 International Conference And Exposition On Electrical And Power Engineering, pp. 591-596, 2012, doi: 10.1109/ICEPE.2012.6463870.

[11] J. Rousselle, J. Blascovich and R. Kelsey, “Cardiorespiratory response under combined psychological and exercise stress,” International Journal Of Psychophysiology, vol. 20, pp. 49-58, 1995.

[12] Y. Can, N. Chalabianloo, D. Ekiz and C. Ersoy, “Continuous Stress Detection Using Wearable Sensors in Real Life: Algorithmic Programming Contest Case Study,” Sensors, vol. 19, p. 1849, Aug. 2019.

[13] R. Costin, C. Rotariu and A. Pasarica, “Mental stress detection using heart rate variability and morphologic variability of EeG signals,” 2012 International Conference And Exposition On Electrical And Power Engineering, pp. 591-596, 2012.

[14] R. Garg, J. Santhosh, A. Dengel and S. Ishimaru, “Stress detection by machine learning and wearable sensors,” The 26th International Conference On Intelligent User Interfaces-Companion, pp. 43-45, April 2021, doi: 10.1145/3397482.3450732.

[15] K. Dalmeida and G. Masala, “HRV features as viable physiological markers for stress detection using wearable devices,” in Sensors, vol. 21, p. 2873, 2021, doi: 10.3390/s21082873.

[16] R. Castaldo, L. Montesinos, P. Melillo, C. James and L. Pecchia, “Ultra-short term HRV features as surrogates of short term HRV: A case study on mental stress detection in real life,” BMC Medical Informatics And Decision Making, vol. 19, pp. 1-13, 2019.

[17] P. Gomes, P. Margaritoff and H. Silva, “pyHRV: Development and evaluation of an open-source python toolbox for heart rate variability (HRV),” Proc. Int’l Conf. On Electrical, Electronic And Computing Engineering (IcETRAN),” pp. 822-828, 2019.

[18] T. Iqbal, A. Simpkin, D. Roshan, N. Glynn, J. Killilea, J. Walsh, G. Molloy, S. Ganly, H. Ryman, E. Coen, A. Elahi, W. Wijns and A. Shahzad, “Stress Monitoring Using Wearable Sensors: A Pilot Study and Stress-Predict Dataset,” Sensors, vol. 22, p. 8135, 2022.

[19] W. Rosenberg, M. Hoting and D. Mandic, “A physiology based model of heart rate variability,” Biomedical Engineering Letters, vol. 9, pp. 425-434, 2019, doi: 10.1007/s13534-019-00124-w.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, “Scikit-learn: Machine Learning in Python,” Journal Of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

[21] J. Montiel, M. Halford, S. Mastelini, G. Bolmier, R. Sourty, R. Vaysse, A. Zouitine, H. Gomes, J. Read, T. Abdessalem, A. Bifet, “River: machine learning for streaming data in Python,” The Journal of Machine Learning Research, vol. 22.1, pp. 4945-4952, 2021.

[22] B. Lakshminarayanan, D. Roy and Y. Teh, “Mondrian forests: Efficient online random forests,” Advances In Neural Information Processing Systems, vol. 27, 2014, pp. 3140-3148, doi: 10.48550/arXiv.1406.2673.

[23] J. Healey and R. Picard, “Detecting stress during real-world driving tasks using physiological sensors,” IEEE Transactions On Intelligent Transportation Systems, vol. 6, pp. 156-166, 2005.

[24] S. Ishaque, N. Khan. and S. Krishnan, “Trends in heart-rate variability signal analysis,” Frontiers In Digital Health, vol. 3, p. 639444, Feb. 2021.