

# Personalized multiclass stress and cognitive load detection

Jaakko Tervonen<sup>1</sup>

<sup>1</sup>VTT Technical Research Centre of Finland, Tekniikantie 1, Espoo, Finland

## Abstract

Stress and cognitive load detection have focused on a binary set-up, where stress is compared to rest, and high cognitive load to lower one. A more detailed analysis could reveal the type of stress, or the moments when a person is approaching high cognitive load, rather than reached it already. In addition, the modelling efforts have focused on finding the best classification algorithm and the biosignals to measure, and other aspects of modelling like the duration of the feature windows, personalization, and model explanations have attained less attention. This study allows stress to have different types and cognitive load to have different levels or to vary continuously. Machine learning methods are investigated to assess the effects of various modelling options in this setting, different signals and features are used to find the best input data, and the influence of the features on the results are discussed. Eye metrics are given special attention since they have been less studied. The study also examines how to personalize the models with little data. The research could make stress and cognitive load detection more precise and widespread e.g. in health domain, education, and safety critical operations.

## Keywords

stress, cognitive load, machine learning

## 1. Context

Stress and cognitive load are complex physiological states that have both positive and negative effects on humans. Although certain level of stress and cognitive load may improve performance [1, 2], cognitive under- and overload may decrease it [2], elevated levels of acute stress can impede performance on cognitive tasks [3], and stress, once chronic, increases risks of e.g. mental health problems and cardiovascular disorders [4]. As both states cause changes to human physiology, developing machine learning (ML) methods to detect the states from wearable sensor measurements has gained significant attention to enable monitoring and interventions.

The ML pipeline in state detection starts with data collection, followed by preprocessing, feature extraction and selection, and classification [5]. Due to individual differences in physiological functions, personalization may be included within one or more of the steps or constitute a step of its own. Whereas the choice of the classification algorithm is frequently experimented with [5], other ML components, such as hyperparameter optimization within classification or feature window duration within preprocessing, are less investigated. They are commonly conducted but results are generally reported only for the best pipeline, hurdling the evaluation of the necessity and effect of the different components affecting classification performance. Despite the advances of deep learning in many fields, it has performed rather poorly in affect recognition and comparably to feature-based classifiers in stress detection [6], which is why a feature-based approach is adopted for this study.

Despite that stress manifests itself in the human body through two stress-responsive axes [7] and thus various types of stress exist, most existing works aim for binary classification between stress and non-stress states [8]. Similarly, despite that cognitive load is predominantly measured with self-report questionnaires working on a continuous scale such as the NASA-TLX [9], attempts to detect it from biosignals are mostly limited to binary classification based on task difficulty, performance, or binarized self-report [10, 11]. The current work aims for multiclass state detection with two stress types (e.g. social and mental stress) and a baseline class or three levels of cognitive load (e.g. low-medium-high).

---

Discovery Science - Doctoral Consortium Contributions 2024

✉ jaakko.tervonen@vtt.fi (J. Tervonen)

ORCID 0000-0003-2236-0253 (J. Tervonen)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Cognitive load may also be modelled as a continuous variable. However, the interrelation of stress and cognitive load is left for future studies.

Both stress and cognitive load can be detected from facial images [12, 13] and functional near-infrared spectroscopy [14] but due to the advent of lightweight wearables measuring several biosignals, focus has been on more unobtrusive and privacy preserving options. To classify between the different states, existing studies have employed various biosignals, such as electrodermal and cardiac activity, skin temperature, and respiration [15, 16]. Although different signal or measurement device combinations are often reported (e.g. [17, 11]), the exact role and the effect of each single feature in state detection are less discussed: model behavior remains unexplained. Cognitive load and stress have also been found to change some eye-related parameters like pupil diameter, saccades, fixations, and blinks [9, 15, 16] but eye measurements are scarcely considered when detecting either stressful or cognitively demanding states.

Baseline physiology and reactions to external stimuli exhibit individual differences which hinders the generalizability of the ML model to new individuals. This is why models are often personalized to account for the differences. The personalization techniques range from individual feature normalization to dividing persons into clusters of similar people and using totally personal models [18]. However, existing methods require a large set of data from each person which is not available in a cold-start, i.e. when a new person starts using the system.

## 2. Research challenges

Existing research on stress and cognitive load detection focuses on binary classification. The studies have used a large number of biosignals but eye measurements have been included less frequently. The exact effect of different features on model behavior and of different ML components affecting performance are little discussed. Furthermore, personalization has focused on techniques requiring extensive data from each end user. This PhD thesis aims to address these shortcomings. The objectives for the thesis are posed as follows:

- O1 To develop a machine learning pipeline to detect multiclass stress and cognitive load, which is explainable and psychophysiological sound.
- O2 To evaluate the contribution of novel eye measurements to stress and cognitive load detection performance.
- O3 To optimize model performance using personalization with minimal data.

## 3. Approach and evaluation

Explorative analyses with various ML pipelines are conducted to develop a modelling approach for O1. The focus is on the contribution of signal segmentation, hyperparameter optimization, classifier and feature selection, and signals to use. Two articles [19, 20] primarily contribute to this objective, including also analyses of model behavior with different methods of explanation. Relating to O2, the latter also includes a wide range of eye parameters to determine their contribution to classification and which eye parameters are most indicative of different stress types.

These two articles assume that all data from each user can be used for personalization. Building on the modelling findings in these articles, two articles [21, 22] evaluate baseline calibration as a means to reduce the amount of data needed for personalization in, respectively, multiclass stress/affect detection and continuous cognitive load detection, with the latter including eye measurements and an explainability analysis of which parameters were most important and how they contributed in the model.

These four articles have now touched each objective but to conclude, one final article will focus more on personalization with the goal of developing methodology for more efficiently personalizing the detection with even less data. Simultaneously, the earlier results regarding the ML pipeline and

the contribution of different signals, especially the eye movements, are evaluated more diversely with several open datasets.

## 4. Preliminary results

The results from the articles already published are summarized as follows:

- [19]: Detecting cognitive load even as a binary variable was rather difficult with a wrist-worn wearable, with classification accuracy of 67.6 % at best. Longer signal segments up to 25 s performed better than shorter ones, with statistical significance. Hyperparameter optimization improved performance by around 2 – 3 %-points on average. Heart rate variability features were the most important.
- [20]: Two types of stress could be distinguished from baseline with up to 86.5 % balanced accuracy with a lab-grade wearable device. Selection of measured signals and selection of features affected the performance more than other modelling choices. Eye movements and cardiac activity were the most important features. The model paid attention to feature changes that are physiologically relevant to the two stress types.
- [22, 21]: A few minutes of baseline data provides sufficient information to calibrate stress/affect and cognitive load detection models personally, outperforming a non-personalized model. However, the requirement for some data remained and the performance with the model that did not limit the amount of data used for personalization was clearly better.

The final article verifying these findings with several open datasets and looking at personalization more closely is expected to be submitted during 2024.

## 5. Discussion and future work

Currently, it seems that characteristics of the used dataset (available signals, measurement device(s), and that tasks participants conducted) affect the modelling performance more than choices made during data analysis. The differences observed between different options related to the modelling pipeline have been rather small. This impression will be elaborated with some more datasets.

The two studies including eye movements have both evidenced that they provide useful, complementary information to the other biosignals and some eye-related features have been among the most important. Thus, future studies would likely benefit from measuring the eyes in addition to the more traditional stress markers.

Lastly, although the benefit of baseline calibration was demonstrated in two articles, the used personalization techniques have so far been related to person-specific feature normalization. More advanced methodology might help to improve the benefit of personalization and further decrease the amount of data needed for it. The remaining article will explore alternative methods for personalization along with a wider analysis of these impressions with several more datasets to generalize the findings.

## Acknowledgments

The PhD work is supervised by Principal Investigator, Docent Jani Mäntyjärvi (VTT Technical Research Centre of Finland). The PhD work is funded by Academy of Finland project 334092, Business Finland project called "Human-technology interoperability and artificial emotional intelligence", and VTT.

## References

- [1] F. S. Dhabhar, The short-term stress response – mother nature’s mechanism for enhancing protection and performance under conditions of threat, challenge, and opportunity, *Frontiers in*

- Neuroendocrinology 49 (2018) 175–192. doi:<https://doi.org/10.1016/j.yfrne.2018.03.004>, stress and the Brain.
- [2] F. Paas, A. Renkl, J. Sweller, Cognitive load theory: Instructional implications of the interaction between information structures and cognitive architecture, *Instructional Science* 32 (2004) 1–8. doi:[10.1023/B:TRUC.0000021806.17516.d0](https://doi.org/10.1023/B:TRUC.0000021806.17516.d0).
  - [3] V. R. LeBlanc, The effects of acute stress on performance: Implications for health professions education, *Academic Medicine* 84 (2009). doi:[10.1097/ACM.0b013e3181b37b8f](https://doi.org/10.1097/ACM.0b013e3181b37b8f).
  - [4] E. A. for Safety and Health at Work, K. Van den Broek, J. Hassard, D. Flemming, R. Gründler, P. Dewe, K. Teoh, B. Cosemans, M. Cosmar, T. Cox, Calculating the costs of work-related stress and psychosocial risks – Literature review, Publications Office, 2014. doi:[doi/10.2802/20493](https://doi.org/10.2802/20493).
  - [5] P. Schmidt, A. Reiss, R. Dürichen, K. V. Laerhoven, Wearable-based affect recognition—a review, *Sensors (Switzerland)* 19 (2019). doi:[10.3390/s19194079](https://doi.org/10.3390/s19194079).
  - [6] M. Dzieżyc, M. Gjoreski, P. Kazienko, S. Saganowski, M. Gams, Can we ditch feature engineering? end-to-end deep learning for affect recognition from physiological sensor data, *Sensors* 20 (2020). doi:[10.3390/s20226535](https://doi.org/10.3390/s20226535).
  - [7] C. F. Sharpley, Neurobiological pathways between chronic stress and depression: Dysregulated adaptive mechanisms?, *Clinical Medicine Insights: Psychiatry* 2 (2009). doi:[10.4137/CMPsy.S3658](https://doi.org/10.4137/CMPsy.S3658).
  - [8] G. Vos, K. Trinh, Z. Sarnyai, M. Rahimi Azghadi, Generalizable machine learning for stress monitoring from wearable devices: A systematic literature review, *International Journal of Medical Informatics* 173 (2023) 221–232. doi:<https://doi.org/10.1016/j.ijmedinf.2023.105026>.
  - [9] T. Kosch, J. Karolus, J. Zagermann, H. Reiterer, A. Schmidt, P. W. Woźniak, A survey on measuring cognitive workload in human-computer interaction, *ACM Comput. Surv.* 55 (2023). doi:[10.1145/3582272](https://doi.org/10.1145/3582272).
  - [10] M. Gjoreski, T. Kolenik, T. Knez, M. Luštrek, M. Gams, H. Gjoreski, V. Pejović, Datasets for cognitive load inference using wearable sensors and psychological traits, *Applied Sciences* 10 (2020). doi:[10.3390/app10113843](https://doi.org/10.3390/app10113843).
  - [11] M. P. Oppelt, A. Foltyn, J. Deuschel, N. R. Lang, N. Holzer, B. M. Eskofier, S. H. Yang, Adabase: A multimodal dataset for cognitive load estimation, *Sensors* 23 (2023). doi:[10.3390/s23010340](https://doi.org/10.3390/s23010340).
  - [12] A. Yüce, H. Gao, G. L. Cuendet, J.-P. Thiran, Action units and their cross-correlations for prediction of cognitive load during driving, *IEEE Transactions on Affective Computing* 8 (2017) 161–175. doi:[10.1109/TAFFC.2016.2584042](https://doi.org/10.1109/TAFFC.2016.2584042).
  - [13] G. Giannakakis, M. R. Koujan, A. Roussos, K. Marias, Automatic stress analysis from facial videos based on deep facial action units recognition, *Pattern Analysis and Applications* 25 (2022) 521–535. doi:[10.1007/s10044-021-01012-9](https://doi.org/10.1007/s10044-021-01012-9).
  - [14] M. Huang, X. Zhang, X. Chen, Y. Mai, X. Wu, J. Zhao, Q. Feng, Joint-channel-connectivity-based feature selection and classification on fnirs for stress detection in decision-making, *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 30 (2022) 1858–1869. doi:[10.1109/TNSRE.2022.3188560](https://doi.org/10.1109/TNSRE.2022.3188560).
  - [15] P. Ayres, J. Y. Lee, F. Paas, J. J. G. van Merriënboer, The validity of physiological measures to identify differences in intrinsic cognitive load, *Frontiers in Psychology* 12 (2021). doi:[10.3389/fpsyg.2021.702538](https://doi.org/10.3389/fpsyg.2021.702538).
  - [16] G. Giannakakis, D. Grigoriadis, K. Giannakaki, O. Simantiraki, A. Roniotis, M. Tsiknakis, Review on psychological stress detection using biosignals, *IEEE Transactions on Affective Computing* 13 (2022) 440–460. doi:[10.1109/TAFFC.2019.2927337](https://doi.org/10.1109/TAFFC.2019.2927337).
  - [17] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, K. Van Laerhoven, Introducing wesad, a multimodal dataset for wearable stress and affect detection, in: *Proceedings of the 20th ACM International Conference on Multimodal Interaction, ICMI '18*, Association for Computing Machinery, New York, NY, USA, 2018, p. 400–408. doi:[10.1145/3242969.3242985](https://doi.org/10.1145/3242969.3242985).
  - [18] J. Tervonen, S. Puttonen, M. J. Sillanpää, L. Hopsu, Z. Homorodi, J. Keränen, J. Pajukanta, A. Tolonen, A. Lämsä, J. Mäntyjärvi, Personalized mental stress detection with self-organizing map: From

laboratory to the field, *Computers in Biology and Medicine* 124 (2020) 103935. doi:<https://doi.org/10.1016/j.combiomed.2020.103935>.

- [19] J. Tervonen, K. Pettersson, J. Mäntyjärvi, Ultra-short window length and feature importance analysis for cognitive load detection from wearable sensors, *Electronics* 10 (2021). doi:[10.3390/electronics10050613](https://doi.org/10.3390/electronics10050613).
- [20] J. Tervonen, J. Närväinen, J. Mäntyjärvi, K. Pettersson, Explainable stress type classification captures physiologically relevant responses in the maastricht acute stress test, *Frontiers in Neuroergonomics* 4 (2023). doi:[10.3389/fnrgo.2023.1294286](https://doi.org/10.3389/fnrgo.2023.1294286).
- [21] J. Tervonen, R. K. Nath, K. Pettersson, J. Närväinen, J. Mäntyjärvi, Cold-start model adaptation: Evaluation of short baseline calibration, in: *Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2023 ACM International Symposium on Wearable Computers*, ACM, 2023. doi:[10.1145/3594739.3610731](https://doi.org/10.1145/3594739.3610731).
- [22] J. Tervonen, R. K. Nath, K. Pettersson, J. Närväinen, J. Mäntyjärvi, Baseline user calibration for cold-start model personalization in mental state estimation, in: *Pervasive Computing Technologies for Healthcare. PH 2023.*, Springer. In Press., Germany, 2023.