Baseline User Calibration for Cold-Start Model Personalization in Mental State Estimation

Jaakko Tervonen[1(✉)], Rajdeep Kumar Nath[2], Kati Pettersson[1], Johanna Närväinen[2], and Jani Mäntyjärvi[3]

[1] VTT Technical Research Centre of Finland, Tekniikantie 1, Espoo, Finland
{jaakko.tervonen,kati.pettersson}@vtt.fi
[2] VTT Technical Research Centre of Finland, Microkatu 1, Kuopio, Finland
{rajdeep.nath,johanna.narvainen}@vtt.fi
[3] VTT Technical Research Centre of Finland, Kaitoväylä 1, Oulu, Finland
jani.mantyjarvi@vtt.fi

Abstract. Robust human state detection based on analysis of physiological signals requires model personalization since physiological reactions are individual. Personalization requires prior information, which is not available for a new, unknown person, i.e. in a cold-start. To overcome this, the current study proposes user calibration, which uses easily obtainable short baseline measurements to normalize physiological variables individually. Experiments were conducted on a cognitive load detection use case to determine effectiveness of the approach, required baseline duration, and the most suitable normalization function. In addition, the behavior of the model was analyzed with Shapley additive explanations to assess its trustworthiness. The results showed that user calibration always beat the non-personalized model, the optimal baseline duration was 3–3.5 min, and there were no differences between the different normalization functions. The model paid the greatest attention to the physiological phenomena found to be indicative of cognitive load in previous studies. The results encourage further evaluation of user calibration in different use cases for smart healthcare.

Keywords: cold-start · physiology · cognitive load · personalization

1 Introduction

Recent advances in sensor technology have enabled pervasive monitoring of people's physiology, which facilitates real-time detection of stress, and mental and cognitive state of the user, to name a few. Knowledge of the user's state can be utilized in the design and implementation of novel interactive applications

The work was funded by VTT and the Academy of Finland under GrantNos: 334092, 351282, 355693.

D. Salvi et al. (Eds.): PH 2023, LNICST 572, pp. 34–48, 2024.
https://doi.org/10.1007/978-3-031-59717-6_3

aiming to improve user health, wellbeing, and performance. For example, a virtual physiotherapist could detect person's activities to assist in physical rehabilitation, detected stress or emotional state could trigger interaction in mental coaching or when recovering from a trauma, measuring alertness or drowsiness with interventions could help if a person has trouble sleeping, and the detected states could be used in clinical decision support as additional information.

To unlock the full potential of these applications, they should work automatically, close to real-time, and adapt to each individual, even new, unknown ones. One major drawback in current state detection approaches is that they fail to properly account for individual differences especially in a cold-start scenario. Basic physiology, reactions to external stimuli and perceptions of varying situations are individual-specific, which should be accounted for in the modelling procedure: the detection model should be personalized. Typically, the physiological features used for state detection are normalized participant-wise, using a whole dataset from each person to do so. This solution accounts for the differences in individual baselines and individual reactions to the different stimuli. However, applying it requires a complete set of data from each participant before the developed model can be applied for them. When a new user starts using the system, i.e. a cold-start occurs, completing a lengthy calibration protocol with different stimuli is burdening for them and may lead to demotivation and even giving up with the system before even properly beginning.

The current study investigates the cold-start problem in the context of cognitive load detection. Monitoring cognitive load is important in safety-critical fields such as flight control, and healthcare professionals working e.g. in the emergency room or the first aid unit, but also in everyday life like driving a vehicle, and in training and education applications for improved learning. It may also help in detecting early signs of cognitive impairments. Furthermore, it has been suggested that cognitive load of medical professionals should be monitored when considering the use of artificial intelligence assisted decision making tools in healthcare [8].

Specifically, using a few minutes of baseline data for model personalization is proposed. Different normalization functions and baseline durations are investigated, and self-reported cognitive load is detected as a continuous variable with a regression model. Additionally, model behavior is explained with a feature contribution analysis with SHAP values. The approach is evaluated on an open-source dataset ADABase [23] having several physiological signals measured in a controlled laboratory protocol consisting of simulated driving and n-back tasks. Such a dataset offers clear signals and tasks which likely results with a rather high cognitive load, making it suitable for the first evaluation of the proposed approach. The main contributions of the study are listed as follows:

- Different normalization strategies are evaluated to use short baseline period for cold-start model personalization in detecting continuous cognitive load.
- Minimal baseline duration for optimal performance is estimated.
- Feature importance and contribution analysis is provided to assess which factors increase and decrease experienced cognitive load.

2 Related Work

2.1 Methods in Cognitive Load Detection

Cognitive load or mental workload refers to the amount of mental resources used to perform a task [24]. Several approaches to measure cognitive load as a continuous variable exist, like self-report questionnaires (e.g. the NASA-TLX [13]), performance measures of cognivitely demanding tasks [30], and physiological triggers, since cognitive load is reflected to e.g. pupillary responses [36,38], heart rate variability [7,28], electrodermal activity [34,35], and facial expressions [16,41]. The link between cognitive load and physiology has led to the development of automated tools to detect cognitive load based on (wearable) sensor data. Figure 1 shows a general machine learning pipeline for the detection task based on sensor data processing.

Previous works attempting cognitive load detection with machine learning methods have primarily focused on a classification setup. Most works have detected high cognitive load from low or no load [3,9,32,37] while some have had three or more levels based on estimated difficulty of the task [18,19]. Although cognitive load can be considered a continuous measure and treating it as a continuous variable allows for a more fine grained analysis, few works attempt to detect it with a regression model. Herbig *et al.* [14,15] have recognized cognitive load in an e-learning and a machine translation task. Pejović *et al.* [25] developed a non-contact sensor to detect cognitive load during elementary cognitive tasks. Lastly, Oppelt *et al.* [23], who introduced the dataset used for experiments in this study, presented results also for regression modelling. Each work selected self-reported cognitive load as the regression target.

2.2 Cold-Start Model Personalization

Baseline physiology, physiological reactions and task perception are individual, which calls for model personalization when detecting cognitive load. Still, several previous studies aim for fully person-independent detection [3,9,19]. When personalization was considered, the most prevalent approach has been some version of participant-wise feature normalization, used in e.g. [14,15,25,32], which normalizes features separately for each person using their full dataset. In addition, it was the only personalization approach considered in a contest to detect cognitive load from wearable sensor data, applied by 5 teams from 12 [11]. Other

Fig. 1. A machine learning pipeline generally used for training a state detection model. Personalization is not always included, which is depicted with a dashed box and it can overlap with either preprocessing or model training.

personalization approaches include e.g. custom domain adaptation [17], where a transfer learning approach adapts the neural network to each individual. In a related context of stress detection, users have been clustered first to train the model based on similar users' data [39].

Each of these approaches require a substantial amount of data from each user, most of them a full set similar to that of the users in the training data. In a cold-start case, such data is not available, and the developed model is unapplicable for new users.

To the best of our knowledge, no earlier studies exist in cognitive load detection addressing this challenge. However, in emotion recognition, Saganowski *et al.* [27] proposed to apply transfer learning with accumulating data in real-life scenario, and in stress detection, participant-wise feature normalization has been implemented with just the baseline data [1,26]. A similar approach for stress and affect detection was examined in [31] who also analyzed the duration of baseline measurement needed. In neuroscientific research, eliminating the individual variations with baseline data is a standard procedure [20].

To set the current study apart from related work the following differences are outlined: i) continuous cognitive load is detected with a regression model as opposed to classification, ii) cold-start is addressed by personalizing based on short baseline measurement, iii) different baseline durations and normalization functions are evaluated.

3 Methodology

3.1 User Calibration

In general terms, a normalization function transforms given input data according to some normalization parameters into a representation that is better suited for a machine learning model. The normalization parameters should be computed from the training and applied for the testing data but the same parameters are used for both splits of the dataset.

In participant-wise scaling, the normalization parameters are computed separately for each participant from a full measurement protocol. Instead, the parameters could be computed from a short baseline measurement, called user calibration in this study, and applied for all subsequent data from that person. In a real-life use case, a new person should sit still and relax for a couple of minutes, allowing the collection of baseline data, which is a much less burdening option than completing the whole protocol. Figure 2 highlights the differences in the inference process between regular personalization and user calibration.

Three normalization functions are applied in this study: averaging $X_{avg} = X - mean(X)$, standard scaling $X_{std} = X_{avg}/std(X)$ and min-max transformation $X_{minmax} = \frac{X-min(X)}{max(X)-min(X)}$, for a dataset X. In participant-wise scaling, the normalization parameters (mean, std, min, max) are computed separately for each participant across the whole measurement protocol, and in user calibration they are computed separately for each participant from short baseline measurement of varying duration.

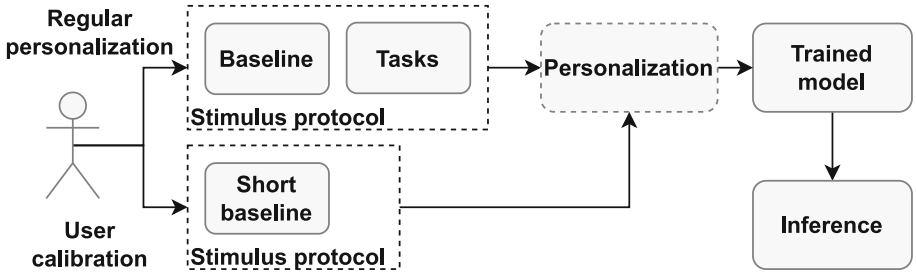

Fig. 2. The cold-start process for mental state estimation of a new user in inference mode. The boxes tagged "Stimulus protocol" display the tasks to be completed before the trained model can be personalized and applied for inferring the state of the user.

3.2 Dataset

The ADABase dataset [23] was adopted for the cognitive load detection experiments. The dataset consists of two tasks aimed at inducing cognitive load: the n-back task and a simulated driving task called k-drive. Simulators and standard cognitive tests provide a controllable and quantifiable environment and thus the induced cognitive states are more homogeneous. Thus, the dataset comprises a good basis to study algorithm development for the cold-start case.

An n-back task consists of a sequence of stimuli and the study participant must indicate when the current stimulus matches the one presented n steps earlier. The n-back task conducted in ADABase consisted of a single (visual stimulus) and a dual stimuli (visual and auditive stimuli) test with three difficulty levels (i.e. $n \in \{1, 2, 3\}$). The k-drive task consisted of watching an autonomous simulator playing a driving game and indicating, on three difficulty levels, whether the car was 1) passing another car, 2) being overtaken, or 3) accelerating or decelerating rapidly; each level was incremental and in each subsequent level the participant had to indicate the events of the previous level(s) as well. Additionally, the participant solved a secondary task of searching and adding songs to a playlist during levels 2 and 3.

The test participants' physiology was monitored with a Biopac MP160 system measuring electrocardiogram (ECG), electromyogram (EMG, trapezius muscle), electrodermal activity (EDA), respiration (RSP), skin temperature (SKT) and photoplethysmogram (PPG). In addition, video-oculography (VOG) was recorded with Tobii Pro Fusion and facial cues with a BASLER camera.

The published version of dataset contains 30 participants, 12 of whom refused the collection of facial video data. The order of n-back and k-drive was randomized, and in the public version 18 participants first completed the n-back task. The measurement set-up was the same for all participants, but it was adjusted for handedness and the time of day varied. The baseline measurement used in this study for user calibration is taken from the resting baseline that occurred before the first stimulus, whether it was n-back or k-drive.

Cognitive load was assessed after each phase of the protocol with the NASA-TLX self-report questionnaire, assessing mental, physical and temporal demand, performance, effort and frustration. Each dimension has a weighting factor to compute the final score as a sum of weighted self-report components. This final score serves as the metric for cognitive load in this study. The authors in the original paper [23] suggest transforming the score individually to a value between 0 and 1 through min-max normalisation. Since this transformation is impossible in a cold-start scenario, it was decided to opt for the unscaled metric in this study.

3.3 Data Processing

Features were extracted from each of the available data sources except for the PPG signal, which was thought redundant since ECG was available. Following the original paper [23], features were extracted with a sliding window of two minutes with a five second window slide. The physiological signals were mostly processed using the NeuroKit2 software package [22], but the heart rate variability features were extracted with the hrv-analysis library [4] and saccades, fixations, and blinks were detected from the VOG data with the PyGaze library [6]. The sum of frames with each facial activation was used as the features for the facial data; see [23] for a description of the activation units. The rest of the extracted features are listed in Table 1.

Table 1. Extracted features from each signal.

Signal	Extracted features
ECG	mean_HR, std_HR, mean_nni, sdnn, nni50, pnni50, rmssd, vlf_power, lf_power, hf_power, lf/hf ratio, total power, lfnu, hfnu, vlf_relative_power, lf_relative_power, hf_relative_power
EMG	rms, n_onsets, fraction_high_activity, max_amplitude
EDA	eda_mean, eda_std, eda_min, eda_max,eda_ slope, eda_range, tonic_mean, tonic_std, tonic_correlation_with_time, phasic_n_peaks, phasic_peak_amplitude, phasic_peak_duration, phasic_peak_area
RSP	breathing_rate_mean, breathing_rate_std, phase_ratio_mean
SKT	skt_mean, skt_std, skt_min, skt_max, skt_slope
VOG	pupil_diam_mean, pupil_diam_std, pupil_diam_slope, blink_rate, blink_duration, time_between_blinks, fix_rate, fix_duration, time_between_fix, n_fix_with_dur_>100ms, n_fix_with_dur_66-150ms, n_fix_with_dur_300-500ms, n_fix_with_dur_>1000ms, sac_rate, sac_duration, time_between_sac, sac_amplitude

ECG = electrocardiogram, EMG = electromyogram, EDA = electrodermal activity, RSP = respiration, SKT = skin temperature, VOG = video-oculography, diam = diameter, fix = fixations, sac = saccades

3.4 Experimental Protocol

The extracted features were used to detect cognitive load as a continuous variable using extreme gradient boosting regressor (XGBoost) [5]. The regressor was selected since extreme gradient boosting has been shown to have good performance in different domains with tabular data, and since it natively handles missing data which is prevalent in the current dataset due to some participants opting out from facial data collection.

The k-drive tasks had a duration of about five minutes and so the experienced cognitive load may have varied over the course of the task. Still, subjective ratings were given only after the task. To ensure that the physiological data is timely and best reflects the given rating, only the last two minute window from each task was used for training the model. This choice has also a balancing effect between the two task types, as the n-back tasks lasted for about two minutes.

Following the original paper [23], the adopted cross-validation strategy was leave-three-users-out, resulting in 10 folds total. For each fold, the data of three randomly selected participants was left out and the model was trained with the remaining participants data and tested on the left out fold. The model performance was assessed in terms of mean absolute error (MAE), mean squared error (MSE) and root mean squared error (RMSE). The performance was compared against a baseline of predicting the average cognitive load for each observation.

Three modelling tasks were defined: i) without feature normalization; ii) with participant-wise feature normalization; and iii) user calibration. The evaluation without feature normalization sets a baseline which user calibration should try and exceed to be useful, and participant-wise feature normalization sets an upper limit of what to expect with normalization-based personalization approaches.

In addition, the duration of the needed baseline measurement was evaluated by training the user calibration model with varying duration of baseline data, ranging from two to five minutes in 30 s increments. Shorter than two minute calibration was not considered since the used feature window length was two minutes. Moreover, three different normalization functions as specified in Sect. 3.1 were experimented with.

The statistical significance of the differences between the personalization and normalization approaches were assessed with related samples T-test, since the cross-validation errors were found to be normally distributed (Shapiro-Wilk test). P-values were corrected with the Benjamini-Hochberg method to adjust the false discovery rate for multiple testing.

4 Results

User calibration performed better than the general and the average prediction model with each normalization function tested (see Table 2) and the difference was statistically significant (Table 3). As expected, participant-wise scaling performed better than user calibration when features were normalized with standard scaling or min-max normalisation, but the two approaches performed the same when features were averaged. No statistically significant differences between the

Table 2. Regression results of predicting subjective cognitive load. Dur stands for best calibration duration in minutes.

Normalization	Personalization	Dur	MAE	MSE	RMSE
Standard scaling	Participant-wise	–	8.20 (1.12)	107.57 (31.86)	10.27 (1.47)
	Calibration	3.0	10.09 (1.69)	148.26 (51.37)	12.01 (2.02)
Averaging	Participant-wise	–	9.93 (1.34)	147.27 (42.49)	12.01 (1.72)
	Calibration	3.5	9.72 (1.54)	136.89 (35.87)	11.60 (1.56)
Min-Max	Participant-wise	–	8.57 (1.75)	114.74 (41.81)	10.52 (2.02)
	Calibration	3.5	9.78 (1.74)	146.99 (47.09)	11.95 (2.05)
No scaling	General model	–	11.03 (1.80)	175.80 (51.65)	13.11 (2.01)
Average prediction	General model	–	10.99 (1.38)	160.78 (34.89)	12.60 (1.42)

Table 3. Related samples T-test results of comparing the MAE's of the user calibration model to those of other models'. P-values were corrected with the Benjamini-Hochberg method.

Normalization	Participant-wise		No scaling		Average prediction	
	T	p	T	p	T	p
Standard scaling	−3.08	0.020	2.49	0.039	−21.64	<0.001
Averaging	0.52	0.618	3.69	0.011	−22.08	<0.001
Min-Max	−2.74	0.029	3.49	0.012	−20.67	<0.001

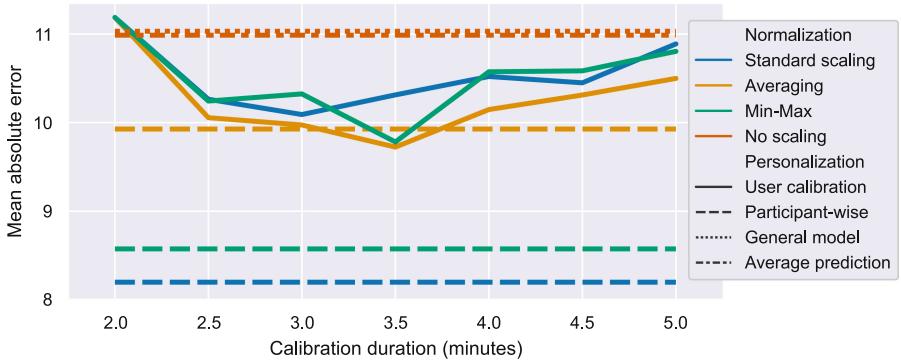

Fig. 3. Model performance at different durations of baseline measurement. The line colors refer to the normalization approach and line style to personalization approach.

three normalization functions when using user calibration were observed (test results not shown). The non-personalized model performed similarly to average prediction.

Figure 3 shows the MAE of user calibration with different baseline durations: a MAE of e.g. 10 would denote that an error of 10 units was made on average

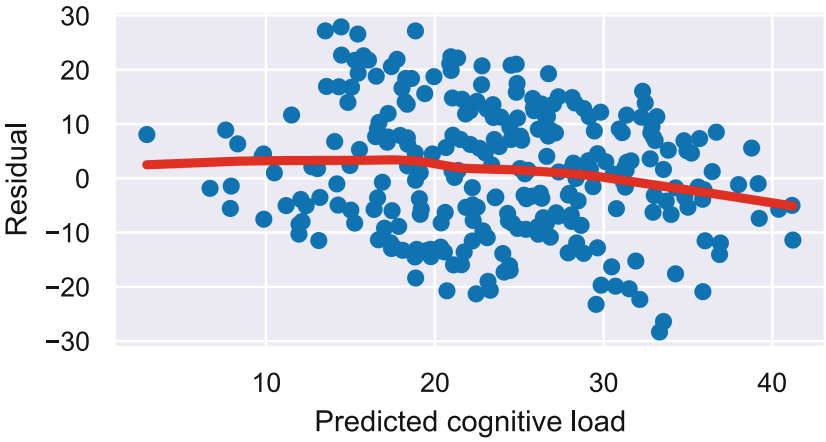

Fig. 4. Scatter plot of model residuals vs. predictions with a LOWESS trend curve, produced over test data folds of the best performing user calibration model with 3.5 min baseline with features normalized with averaging.

when predicting the load. The best performance was observed with 3 min baseline duration with standard scaling, and 3.5 min with averaging and min-max normalization. However, the differences between the different durations were not large.

According to the residual plot in Fig. 4, there were no clear signs of heteroscedasticity or outliers. The trend curve shows slight elevation at lower cognitive loads, and a minor decreasing trend towards the higher predicted cognitive load estimates. Thus, the model may overestimate lower cognitive load and underestimate higher one, but based on the figure the effect should not be large.

Shapley additive explanations (SHAP) [21] were computed to assess the importance of different signal modalities and features. Figure 5 displays a beeswarm plot of the top-20 features with highest average absolute SHAP values in a decreasing order, drawn over the best performing user calibration model. Judging by the number of features from different modalities in the plot, the most influential signal modalities were facial activations and ECG, and each signal modality had at least one feature within the top-20.

Although the direction of changes is a little confused for some features, certain conclusions can be made from the figure. The most influential feature was mean pupil diameter, with its higher values corresponding to higher cognitive load, and vice versa for lower values. Lower/higher values in AU20 (lip stretcher), AU25 (lip part), and AU05 (upper lid raiser) corresponded to lower/higher cognitive load, respectively. Moreover, higher skin temperature, variation in breathing rate, blink duration, trapezoidal EMG activity, and electrodermal activity, and lower heart rate variability all corresponded to higher cognitive load, according to the developed model. These are roughly in line with previous observations [2,7,35,36] but there was mixed evidence of skin temperature and respiration changes under cognitive load [10,12,34].

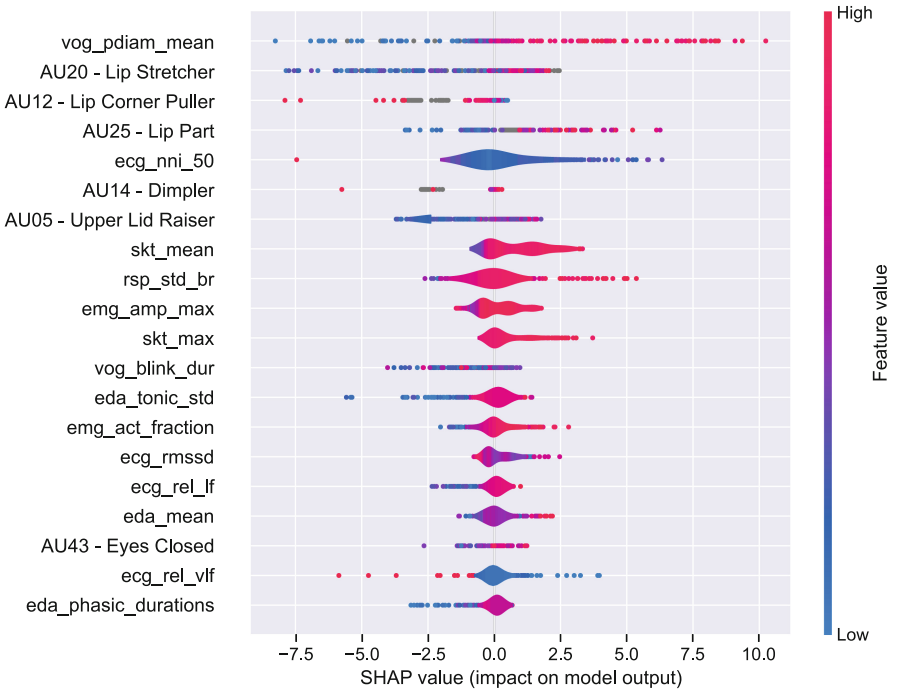

Fig. 5. Beeswarm plot of the SHAP values of the features in the user calibration model with 3.5 min baseline with features normalized with averaging.

5 Discussion

The presented analysis proved the usefulness of user calibration, since it always outperformed the general model regardless of the choice of the normalization function. Thus, collecting 3–3.5 min of baseline resting data from a new user allows making better predictions since the beginning. The proposed approach has some limitations which form the ground for future work: it only captures individual differences at the level of basic physiology, determining the correct ground truth is challenging, the used dataset was collected in a constrained environment with a large set of physiological signals, and only one modelling approach was evaluated. Each of these points are further discussed next.

The data collection protocol exhibited individual variations in three levels: basic physiology, reactions to stimuli, and self-reporting. The basic physiology between the participants differed, which was up to some extent caught by user calibration. Different people react in different ways to similar external stimuli, which in turn was caught by participant-wise normalization. Finally, different people may experience the same task in various ways: some see the task demanding while others find it enjoyable, which is reflected in the subjective reports. Such information could be included in the model by normalizing the reported

load participant-wise, as was done in the original paper [23], but not here since it would be unfeasible in a cold-start scenario.

Curiously, the general model performed the same as predicting the average score. Since the general model has no knowledge of any of the individual components, this further highlights the need for personalization especially in the cold-start case: if there was no prior information from a new individual, using the general, non-personalized model for them would be as good as guessing the average score.

Obtaining this prior information is a crucial step in cold-start model personalization for mental state estimation: without some, one cannot personalize. Having the people rest for a few minutes, as in this study, is one of the less burdensome techniques, but it contains no information about how people might react to or perceive cognitively loading stimuli. While it is already a more tiring option, conducting a mini-protocol of short cognitive tasks might be an alternative to include reactions and variation in the baseline data. However, any baseline measurement should be repeated periodically since e.g. caffeine intake [40] or fatigue [33] may cause changes to human physiology. Therefore, one could alternatively opt for collaborative filtering [29] based methods by e.g. measuring user similarity via background questionnaire, such as demographics or personality traits. Although the big five personality traits did not improve classification results in [23], they could still be useful in cold-start personalization.

Due to subjectiveness, determining the ground truth in detection of cognitive load and other mental states is also a challenge of its own. The current study used non-normalized cognitive load self-reports as labels for a regression model. One could also choose to normalize the labels participant-wise (not feasible in a cold-start scenario) or choose to classify between self-reported low/medium/high load or even classify based on task labels. The latest would be an objective measure when looking at the data labelling, but still physiology would reflect subjective load and some people might not experience cognitive load in tasks labelled under high load. Ultimately the choice of the ground truth comes down to the targeted use case: what is most sensible given the context? The target used in the current study is suitable for applications where the interest is in subtle subjective changes in cognitive load. A continuous target and regression analysis applies to a more (time-wise) continuous modelling and allows developing methods to detect moments when the person is just heading towards cognitive overload, unlike classification, which provides a more definite outcome.

The dataset used in this study contained a simulated driving task and the n-back task conducted in a controlled laboratory environment. Since this was the first inspection of cold-start model personalization in cognitive load detection, such dataset with clear and likely high cognitive load periods was chosen to be able to focus on the cold-start issue. Since the results were encouraging, the method should be evaluated next on different datasets from the health domain to improve its ecological validity.

While some features showed a rather clear behavior pattern in Fig. 5, like pupil diameter and skin temperature, the behavior of other features, such as

RMSSD and maximum amplitude of EMG, was more convoluted. Here, too, there may be some individual differences. Indeed, several features are clustered around zero, meaning that for those observations the current feature had a small impact, and long tail(s) denoting observations for which the current feature had a larger impact. However, the coloring of e.g. RMSSD and EMG_max_amplitude show that higher values were related to both decreased and increased cognitive load. These may be related to some spurious events during the completion of the task, changing of cognitive load during the two-minute window, or the relation between self-reports and some physiological parameters may be unsystematic.

Investigation and understanding this relationship and finding features with systematic behavior under varying cognitive load is a necessity for robust cognitive load detection. To keep focus on user calibration and normalization options, this analysis is left for future work, together with finding the best type of a model, hyperparameter optimization, feature window duration optimization, feature selection, and signal modality selection, all of which are important steps to consider in model development. Based on Fig. 4, the developed model fit to the data reasonably well, but the described steps may help in improving the model performance.

6 Conclusions

Overcoming the cold-start situation in model personalization is a necessity for future human state detection applications. In this study, using short baseline measurements to normalize features was proposed as the solution in detecting continuous cognitive load. The experiments showed that user calibration always performed better than the general model but worse than a model with participant-wise normalization with full dataset. The optimal baseline duration was found to be 3–3.5 min and there were no differences between the tested normalization functions. A SHAP feature importance analysis revealed that the developed model found physiologically correct patterns, increasing trust to it. Future studies are needed for different mental states to further validate the proposed user calibration approach.

References

1. Albaladejo-González, M., Ruipérez-Valiente, J.A., Gómez Mármol, F.: Evaluating different configurations of machine learning models and their transfer learning capabilities for stress detection using heart rate. J. Ambient. Intell. Humaniz. Comput. (2022). https://doi.org/10.1007/s12652-022-04365-z
2. Biondi, F.N., Cacanindin, A., Douglas, C., Cort, J.: Overloaded and at work: investigating the effect of cognitive workload on assembly task performance. Hum. Factors **63**(5), 813–820 (2021). https://doi.org/10.1177/0018720820929928
3. Bozkir, E., Geisler, D., Kasneci, E.: Person independent, privacy preserving, and real time assessment of cognitive load using eye tracking in a virtual reality setup. In: 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), pp. 1834–1837 (2019). https://doi.org/10.1109/VR.2019.8797758

4. Champseix, R.: Heart Rate Variability analysis (2018). https://github.com/Aura-healthcare/hrv-analysis. Accessed 20 June 2023
5. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, vol. 19, pp. 785–794. ACM, New York, NY, USA, August 2016. https://doi.org/10.1145/2939672.2939785, https://dl.acm.org/doi/10.1145/2939672.2939785
6. Dalmaijer, E.S., Mathôt, S., Van der Stigchel, S.: Pygaze: an open-source, cross-platform toolbox for minimal-effort programming of eyetracking experiments. Behav. Res. Methods **46**(4), 913–921 (2014). https://doi.org/10.3758/s13428-013-0422-2
7. Delliaux, S., Delaforge, A., Deharo, J.C., Chaumet, G.: Mental workload alters heart rate variability, lowering non-linear dynamics. Front. Physiol. **10** (2019). https://doi.org/10.3389/fphys.2019.00565
8. Ehrmann, D.E., et al.: Evaluating and reducing cognitive load should be a priority for machine learning in healthcare. Nat. Med. **28**(7), 1331–1333 (2022). https://doi.org/10.1038/s41591-022-01833-z
9. Feradov, F., Ganchev, T., Markova, V.: Automated detection of cognitive load from peripheral physiological signals based on Hjorth's parameters. In: 2020 International Conference on Biomedical Innovations and Applications (BIA), pp. 85–88 (2020). https://doi.org/10.1109/BIA50171.2020.9244287
10. Gjoreski, M., et al.: Datasets for cognitive load inference using wearable sensors and psychological traits. Appl. Sci. **10**(11) (2020). https://doi.org/10.3390/app10113843
11. Gjoreski, M., et al.: Cognitive load monitoring with wearables-lessons learned from a machine learning challenge. IEEE Access **9**, 103325–103336 (2021). https://doi.org/10.1109/ACCESS.2021.3093216
12. Grassmann, M., Vlemincx, E., von Leupoldt, A., Mittelstädt, J.M., Van den Bergh, O.: Respiratory changes in response to cognitive load: a systematic review. Neural Plast. **2016**, 8146809 (2016). https://doi.org/10.1155/2016/8146809
13. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (task load index): results of empirical and theoretical research. In: Hancock, P.A., Meshkati, N. (eds.) Human Mental Workload, Advances in Psychology, vol. 52, pp. 139–183. North-Holland (1988). https://doi.org/10.1016/S0166-4115(08)62386-9
14. Herbig, N., et al.: Investigating multi-modal measures for cognitive load detection in e-learning. In: Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization. UMAP '20, pp. 88–97. Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3340631.3394861
15. Herbig, N., Pal, S., Vela, M., Krüger, A., van Genabith, J.: Multi-modal indicators for estimating perceived cognitive load in post-editing of machine translation. Mach. Transl. **33**(1), 91–115 (2019). https://doi.org/10.1007/s10590-019-09227-8
16. Hussain, M.S., Calvo, R.A., Chen, F.: Automatic cognitive load detection from face, physiology, task performance and fusion during affective interference. Interact. Comput. **26**(3), 256–268 (2013). https://doi.org/10.1093/iwc/iwt032
17. Jiménez-Guarneros, M., Gómez-Gil, P.: Custom domain adaptation: a new method for cross-subject, EEG-based cognitive load recognition. IEEE Sig. Process. Lett. **27**, 750–754 (2020). https://doi.org/10.1109/LSP.2020.2989663
18. Khanam, F., Hossain, A.A., Ahmad, M.: Electroencephalogram-based cognitive load level classification using wavelet decomposition and support vector machine. Brain-Comput. Interfaces **10**(1), 1–15 (2023). https://doi.org/10.1080/2326263X.2022.2109855

19. Li, Y., Li, K., Wang, S., Li, Y., Chen, J., Wen, D.: Towards safer flights: a multimodality fusion technology-based cognitive load recognition framework. In: 2022 IEEE 4th International Conference on Civil Aviation Safety and Information Technology (ICCASIT), pp. 525–530 (2022). https://doi.org/10.1109/ICCASIT55263.2022.9986937
20. Luck, S.J.: An Introduction to the Event-Related Potential Technique. MIT Press, Cambridge (2014)
21. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS'17, pp. 4768–4777. Curran Associates Inc., Red Hook, NY, USA (2017). https://doi.org/10.5555/3295222.3295230
22. Makowski, D., et al.: NeuroKit2: a Python toolbox for neurophysiological signal processing. Behav. Res. Methods **53**(4), 1689–1696 (2021). https://doi.org/10.3758/s13428-020-01516-y
23. Oppelt, M.P., et al.: Adabase: a multimodal dataset for cognitive load estimation. Sensors **23**(1) (2023). https://doi.org/10.3390/s23010340
24. Orru, G., Longo, L.: The evolution of cognitive load theory and the measurement of its intrinsic, extraneous and germane loads: a review. In: Longo, L., Leva, M.C. (eds.) H-WORKLOAD 2018. CCIS, vol. 1012, pp. 23–48. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-14273-5_3
25. Pejović, V., Matkovič, T., Ciglarič, M.: Wireless ranging for contactless cognitive load inference in ubiquitous computing. Int. J. Hum.-Comput. Interact. **37**(19), 1849–1873 (2021). https://doi.org/10.1080/10447318.2021.1913860
26. Prajod, P., André, E.: On the generalizability of ECG-based stress detection models. In: 2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 549–554 (2022). https://doi.org/10.1109/ICMLA55696.2022.00090
27. Saganowski, S., Kunc, D., Perz, B., Komoszyńska, J., Behnke, M., Kazienko, P.: The cold start problem and per-group personalization in real-life emotion recognition with wearables. In: 2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops), pp. 812–817 (2022). https://doi.org/10.1109/PerComWorkshops53856.2022.9767233
28. Solhjoo, S., et al.: Heart rate and heart rate variability correlate with clinical reasoning performance and self-reported measures of cognitive load. Sci. Rep. **9**(1), 14668 (2019). https://doi.org/10.1038/s41598-019-50280-3
29. Su, X., Khoshgoftaar, T.M.: A survey of collaborative filtering techniques. Adv. Artif. Intell. **2009**, 421425 (2009). https://doi.org/10.1155/2009/421425
30. Sweller, J., Ayres, P., Kalyuga, S.: Measuring Cognitive Load, pp. 71–85. Springer, New York (2011). https://doi.org/10.1007/978-1-4419-8126-4_6
31. Tervonen, J., Nath, R.K., Pettersson, K., Närväinen, J., Mäntyjärvi, J.: 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 the 2023 ACM International Symposium on Wearable Computing. UbiComp/ISWC '23 Adjunct, pp. 417–422. Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3594739.3610731
32. Tervonen, J., Pettersson, K., Mäntyjärvi, J.: Ultra-short window length and feature importance analysis for cognitive load detection from wearable sensors. Electronics **10**(5) (2021). https://doi.org/10.3390/electronics10050613

33. Tran, Y., Wijesuriya, N., Tarvainen, M., Karjalainen, P., Craig, A.: The relationship between spectral changes in heart rate variability and fatigue. J. Psychophysiol. **23**(3), 143–151 (2009). https://doi.org/10.1027/0269-8803.23.3.143
34. Vanneste, P., et al.: Towards measuring cognitive load through multimodal physiological data. Cogn. Technol. Work **23**(3), 567–585 (2021). https://doi.org/10.1007/s10111-020-00641-0
35. Visnovcova, Z., Mestanik, M., Gala, M., Mestanikova, A., Tonhajzerova, I.: The complexity of electrodermal activity is altered in mental cognitive stressors. Comput. Biol. Med. **79**, 123–129 (2016). https://doi.org/10.1016/j.compbiomed.2016.10.014
36. Volden, F., De Alwis Edirisinghe, V., Fostervold, K.-I.: Human gaze-parameters as an indicator of mental workload. In: Bagnara, S., Tartaglia, R., Albolino, S., Alexander, T., Fujita, Y. (eds.) IEA 2018. AISC, vol. 827, pp. 209–215. Springer, Cham (2019). https://doi.org/10.1007/978-3-319-96059-3_23
37. Wu, C., Liu, Y., Guo, X., Zhu, T., Bao, Z.: Enhancing the feasibility of cognitive load recognition in remote learning using physiological measures and an adaptive feature recalibration convolutional neural network. Med. Biol. Eng. Comput. **60**(12), 3447–3460 (2022). https://doi.org/10.1007/s11517-022-02670-5
38. Xu, J., Wang, Y., Chen, F., Choi, E.: Pupillary response based cognitive workload measurement under luminance changes. In: Campos, P., Graham, N., Jorge, J., Nunes, N., Palanque, P., Winckler, M. (eds.) INTERACT 2011. LNCS, vol. 6947, pp. 178–185. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23771-3_14
39. Xu, Q., Nwe, T.L., Guan, C.: Cluster-based analysis for personalized stress evaluation using physiological signals. IEEE J. Biomed. Health Inform. **19**(1), 275–281 (2015). https://doi.org/10.1109/JBHI.2014.2311044
40. Yeragani, V.K., Krishnan, S., Engels, H.J., Gretebeck, R.: Effects of caffeine on linear and nonlinear measures of heart rate variability before and after exercise. Depress. Anxiety **21**(3), 130–134 (2005). https://doi.org/10.1002/da.20061
41. Yüce, A., Gao, H., Cuendet, G.L., Thiran, J.P.: Action units and their cross-correlations for prediction of cognitive load during driving. IEEE Trans. Affect. Comput. **8**(2), 161–175 (2017). https://doi.org/10.1109/TAFFC.2016.2584042