Load Induction then Simultaneous Relaxation: Insights from Multi-Modal Time-Series Data Measured with Low-Cost Wearable Sensors

Christoph Anders, Sai Siddhant Gadamsetti, Nico Steckhan, Bert Arnrich
Digital Health - Connected Healthcare
University of Potsdam
Hasso Plattner Institute
14482 Potsdam, Germany
e-mail: Christoph.Anders@hpi.de, Siddhant.Gadamsetti@hpi.de, Nico.Steckhan@hpi.de, Bert.Arnrich@hpi.de

***Abstract*—Prolonged levels of high mental workload and resulting stress are among the main causes of employee sickness. A possible solution would be implementing business rules based on objective analyses of stress levels and cognitive demands produced in employees by given tasks. This study laid the foundation for the development of personalized stress assistants. Physiological data of five groups of two participants were recorded, following a five-appointment study design. During the appointments, each pair underwent a cognitive load induction and subsequent stress reduction phase. Physiological signals were recorded with low-cost wearable sensors, subsequently analyzed for biomarkers, and compared for similarity between participants and groups. Results show that the sensors are capable of capturing descriptive data. Despite simultaneous task executions, it was found from the similarity analysis that the normalized Dynamic Time Warping distances between extracted features are greater for yoga sessions than during the cognitive load sessions. The classification of tasks was performed using the Machine Learning algorithms (i) Logistic Regression, (ii) Support Vector Machines, (iii) Nearest Neighbors, and (iv) Decision Trees trained on feature sets of either the Muse S, the Empatica E4, or both sensors together. Generalized as well as personalized models achieved classification accuracies over 85.00%. The recorded data is available upon request. The stimulus elicitation framework developed using PsychoPy and the software artifacts for data analysis were made publicly available, enabling the research community to evaluate their methods on this dataset and re-use analysis methods on their own or other datasets.**

Keywords*—*Mental Health; Mental Workload; Stress; Wearables; eHealth.

I. INTRODUCTION

To perform any natural task, humans utilize mental resources. In this context, a widely referenced concept is mental workload. According to [1], "*Mental workload may be viewed as the difference between the capacities of the information processing system that are required for task performance to satisfy performance expectations and the capacity available at any given time.*". It has been shown that the risk of coronary heart disease and hypertension, amongst other diseases, is increased if the mental workload is sustained at an elevated level over a long time, as mental workload alters the cardiovascular function, leading to a rising heart rate and blood pressure [2], [3].

To counteract such adverse consequences, these elevated levels of mental workload first need to be identified. Different avenues exist, such as performance-based, subjective, and physiological approaches. Performance-based measures mainly highlight situations where high levels of mental workload lead to mental overload. Subjective measures include self-assessments, but it has been shown that humans perform poorly in self-identifying decreased vigilance and cognitive overload [4]. Physiological measures are based on changes in the body incurred by mental workload, such as pupil dilation, heart rate, and changes in skin conductance. These measures can work on a continuous scale but usually require specialized equipment and trained staff [5].

A review on measuring mental workload covering Electrocardiogram (ECG), blood pressure, respiratory, ocular and dermal sensors alongside Electroencephalography (EEG), found that different measures can be used to discriminate task load, task type, and task difficulty while underlining the importance of multi-modal setups [6]. Furthermore, it was shown that one-channel in-ear EEG might suffice in optimal circumstances [7], while stress reduction can be predicted using ECG data from wearable sensors, amongst others [8]. As for mental workload, another interesting phenomenon was observed: by unconscious synchronization of brain activity across individuals, these individuals might utilize more mental resources than each individual alone would be able to [9]. This phenomenon was studied in various settings, such as communication [10] and learning processes between teachers and students, where the strength of the personal bond was found to be a modulator [11], [12].

To the best of the authors' knowledge, no related work focused on incorporating the analysis of group-wide processes of physiological signals in evaluating mental workload, stress, and stress-reduction interventions. Here, the reliability of wearable sensor systems on mental workload, stress, and activity type classification was investigated. Furthermore, a similarity analysis pipeline using the well-studied oddball paradigm [13] was validated, to quantify the effect of a yoga intervention in reducing mental workload and stress.

The remainder of the work is structured as follows: in Section II, related work is presented, while Section III details the methods employed in this work. Section IV gives the results of this work, for which future work is given in Section V. Finally, the conclusion is given in Section VI.

II. RELATED WORK

It was found that EEG measurements have adequate time resolution, conveying information online, and thus providing a promising tool for assessing cognitive workload, comparable in simplicity to measuring the physical workload with heart rate monitors or pedometers [4]. This finding was extended by another review of mental workload classification using wearable on-body devices, finding that EEG seems most promising and should be included in every multi-modal setup, as 'it is the only method that is directly related to mental workload' according to [14].

Given the negative effect of distress, interventions to reduce the stress levels of participants are plentiful. As such, studies have investigated the effects of exposure to music and nature sounds [15], mind-body connection courses designed to reduce anxiety [16], and multi-dimensional stress reduction interventions employing cognitive, somatic, dynamic, emotive and hands-on interventions [17], amongst many more. In addition, various literature reviews were conducted on this topic [18], [19]. Yoga and breathing exercises are widely known as a specific form of mindfulness, practiced in various forms for thousands of years. Numerous literature reviews synthesized some of the key findings for yoga on individuals concerning reductions in depression symptoms, stress and anxiety ratings, as well as the frequency of symptoms, such as headaches, particularly also in a short time frame after the onset of the intervention [20]–[22]. It was found that practices that include yoga asanas appear to be associated with improved regulation of the sympathetic nervous system and hypothalamic-pituitary-adrenal system [22].

In light of movement-based interventions, the contamination of physiological time series with movement artifacts needs to be considered. As for ocular artifacts (looking at instructive yoga videos in the present study), conflicting evidence was found. One work found that no substantial artifacts were present in mobile EEG readings, naturally except for frontal recording sites [23], and another work found that eye movements significantly distorted recordings from electrodes at frontal, temporal, and ear positions [24]. Both works agree, however, that artifacts are generally stronger in EEG bands of higher frequency. Automatic artifact tagging algorithms were proposed, to classify movement artifacts as emerging from loss of contact with the sensor, or from movement of the underlying tissue, as demonstrated on EEG data [25]. Recently, the current state of the art of movement artifact removal from EEG was summarized, finding that software and hardware solutions need to be utilized simultaneously, and recommending guidelines [26].

As for another modality, the Photoplethysmography (PPG), it was found that wavelet transforms as well as Kalman filters might be needed to remove unwanted artifacts from the data, mitigating the impact of artifacts [27]. With the rise of Machine Learning (ML) techniques, artifact detection has shifted to employ such measures as well, as demonstrated by unsupervised artifact identification in another modality recorded at the wrist: electrodermal activity [28].

While well-studied event elicitation tests exist, (such as the Oddball paradigm, which is widely used for the analysis of event-related potentials in schizophrenia patients [29]), and synchronization algorithms for wearable devices exist (e.g., [30]), measurements of synchronicity of event-related responses recorded with wearable sensors are rarely but effectively performed [13]. The utilization of similarity measures for physiological data has recently gained some attention, especially for clinical decision support systems [31], but has, to the best of the authors' knowledge, rarely been performed for simultaneously recorded physiological data from wearable devices.

III. METHODS

Many challenges come up when working to synchronously record data from multiple participants, potentially even more so with wearable sensors than with hard-wired clinical-grade devices. As experimenters are usually not trained clinicians, the sensor fit of wearable devices is often of poorer quality than any clinical counterpart, with participant movements worsening the signal quality as described. Furthermore, signal transmission is mostly performed via third-party apps without explicit support for synchronous data recordings, shielding the experimenters from working with proprietary communication channels, while hiding a lot of the complexity inherent in synchronous data channels and potentially performing data cleaning on the (asynchronously) recorded data. This can lead to reduced trust in the recorded data if it was wholly recorded synchronously, or if some sensor clock-drift occurred or samples were dropped and interpolated at another time. To enable the research community to perform synchronous recordings in a multi-sensor and multi-user setup, a technical feasibility study was conducted in this work, including the conceptualization, development, and validation of a novel technical recording framework.

A. Utilized Sensors

As wearable sensors, the widely utilized wearable devices Empatica E4 and Muse S were employed. The Empatica E4 is a wrist-worn device, which contains Photoplethysmography (PPG; sensor read-out used to measure changes in the blood volume pulse), Electrodermal Activity (EDA, skin conductance measure correlating to stress, mental workload, and emotional responses), and accelerometer sensors. The Muse S headband contains four Electroencephalogram (EEG; records changes of the brain's electrical activity) sensors placed according to the 10/20 international system. Two frontal electrodes (AF7 and AF8) rest on the forehead and two

ISBN: 978-1-68558-167-1

temporal electrodes (TP9 and TP10) rest behind the ears. A reference sensor is located at the center of the forehead (FpZ). Apart from EEG sensors, Muse S contains PPG, gyroscope, and accelerometer sensors. Both wearables are commercial off-the-shelf devices, which have been tested and certified for safety under various regulatory standards, such as FCC and CE. The data was collected from the devices via a newly developed recording platform, implemented in Python 3.9 and building on top of PyLSL [32], a Python interface to the Lab Streaming Layer (LSL)) as well as on top of the Empatica E4 streaming server for Windows. For each wearable sensor, a separate BLED112 Bluetooth Dongle had to be utilized. The source code of the recording framework has been made publicly available at [33].

B. Study Design

During the study, five groups of two participants underwent five recording sessions on individual days. Each recording session lasted approximately 90 minutes, split into welcoming the respective pair of participants and fitting the sensors, a stress induction phase of approximately 30 minutes, and a yoga intervention of approximately 30 minutes succeeding the phase of high mental workload. Before, in between, and after the activities, subjective questionnaire data was collected from the participants. However, the yoga practice has not been interrupted to collect questionnaire answers, and as such subjective mental state assessments were collected only before and after the yoga practice. As questionnaires, the Brunel Mood Scale Questionnaire (BRUMS-Q), Stanford Sleepiness Scale (SSS), Visual Analogue Scale to Evaluate Fatigue Severity (VAS-F), as well as five-point Likert scales in the dimensions of mental workload and stress were utilized.

The induction of mental workload and stress was realized using randomized assignments of the widely used mental workload tasks AX-Continuous Performance Task (AX-CPT) [34] and Time Load Dual Back Task (TloadDback) [35], implemented in Python and presented using the PsychoPy platform [36]. Figure 2 gives an overview of the cognitive load induction framework. For the intervention, a publicly available Yoga video [37] was reproduced on a 75-inch TV screen. Half of the recordings (12 sessions) took place in a controlled environment at the Hasso Plattner Institute Campus 3, House G2, in Potsdam, Germany, a well-illuminated room with floor-to-ceiling windows on two sides of the room offering a view to trees. The other half of the recordings took place in uncontrolled environments. Out of a variety of options, the homes of some of the participants were chosen as uncontrolled environments at the request of the participants. Repeating some yoga poses, a sequence of 29 asanas was performed and finished with Shavasana and a chant of Om. Figure 1 gives a schematic overview of the study design. The cognitive load induction is described in detail in Figure 2. After twenty trials, the performance was assessed. If less than 85% of correct responses were achieved, the system added 0.1 seconds to each Stimulus Time (ST) and Response Time (RT) and repeated the process of Individualization. However, if the user had achieved 85% performance or more, the framework moved on with the current ST and RT settings to the final task for the remaining duration of the cognitive load induction phase.

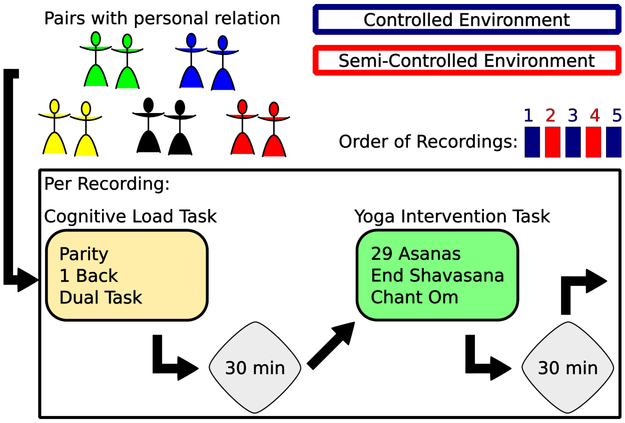

Figure 1: Overview of the study design. As for the Yoga intervention, 20 unique asanas were utilized by the video instructor (e.g., Child Pose, Cat and Cow, Downward-Facing Dog, etc.).

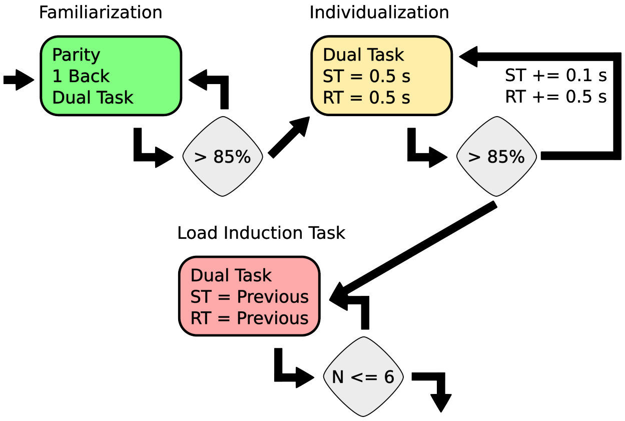

Figure 2: Overview of the cognitive load induction framework. Participants were first familiarized with the individual tasks. After a total of 60 trials, the participant's performance was assessed. If more than 85% of the cues were responded correctly, the user moved on to the individualization phase, which started directly with the lowest Stimulus Time (ST) and Response Time (RT).

Ethical approval has been obtained from the Institutional Review Board (IRB) of the University of Potsdam (application number 69/2023), and written informed consent was given by all participants before participating in the study. The study inclusion criteria required participants to be aged 18 to 33, sufficiently fluent in English (at least B2 level), have a normal or corrected-to-normal vision, know how to use a smartphone, and have to regularly perform work that was performance-evaluated (e.g., students or employees). Participants were required to regularly perform sports or yoga, to be experienced with moderate at-home workouts, stretching, and video-based yoga, and to be in a close relationship with the participant they registered with.

The study exclusion criteria excluded participants who needed to regularly take medication, such as mood stabilizers or psychotropic drugs and could not record data for approximately 90 minutes without interruptions except for bathroom visits.

As the study required the participants to perform specific yoga exercises, physically disabled or injured persons (recovered for less than six months) had to be excluded in case the prospective participant was unable to perform the majority of the required movements. Furthermore, participants who could have been in any dependent relationship with the experimenters, pregnant women, and participants with hypertension were excluded. Out of an overwhelming response to study recruitment efforts, a random total number of ten participants were recruited and recorded to evaluate the technical feasibility of the study setup.

C. Similarity

To confirm the synchrony of the recorded data, two sanity checks were integrated into the study protocol. Firstly, the experimenters vigorously shook the recording devices at the beginning and end of the recordings for approximately ten seconds. This ensured simultaneous peaks in the acceleration data of the wearables, and as a result enabled the comparison of peak onset and offset times, validating that no clock drift had occurred during the recordings and that by consequence the time series between the well-aligned peaks in the start and at the end of the recordings had to be well-aligned as well. Secondly, an Oddball paradigm was utilized to validate if it was possible to measure Steady-State Visual Evoked Potentials (SSVEPs) with the Muse S wearable EEG headband and to analyze the synchrony of these SSVEPs.

However, due to calibration issues with the TV screen, the majority of the Oddball paradigm sessions were not reproduced with the anticipated 60 Hz refresh rate of the screen and a matching signal rate, but with a refresh rate much lower, resulting in invalid Oddball recordings that had to be interrupted due to excessive durations and very slow signal changes. Due to Bluetooth data transmission and Bluetooth channel saturation, drops in sampling frequencies of the individual sensors occurred. Mostly, however, the Muse S sampled EEG data at 256 Hz, PPG data at 64 Hz, and Gyroscope and Acceleration data at 50 Hz. The Empatica E4 mostly sampled BVP data at 64 Hz, Acceleration data at 32 Hz, and GSR as well as Skin Temperature data at 4 Hz.

D. Data Processing

During data recording, the data was stored in .h5 format. After each recording session was stopped, the newly developed streaming platform StreamSense immediately triggered a data cleaning and data processing pipeline ProSense, creating signal quality reports and subsequently storing the recorded data in .pkl format. Figure 3 gives an abstract representation of the data preprocessing flow, triggered automatically after each recording session. The individual parameters, such as outlier rejection thresholds for the dynamic Interquartile Range (IQR) method, pass- and stop-band definitions, as well as the normalization method utilized (min-max), are documented in the source code documentation of ProSense. Alongside the sensor data, log files were created from Questionnaire answers, performance times, reaction times, and system logs. For each recording, the logs were cleaned, a processed subset was stored, and features were extracted and stored in individual .csv files corresponding to the respective modality.

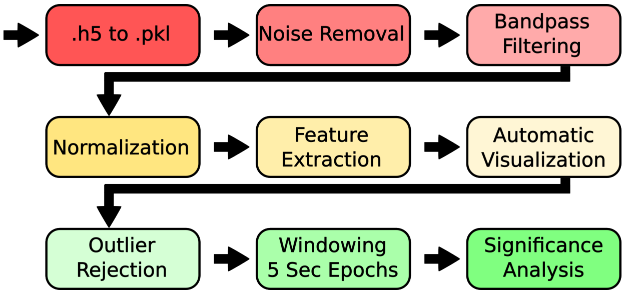

Figure 3: Overview of the data processing pipeline, triggered automatically by ProSense after each recording.

Across files, the same (anonymized) identifiers for participants as well as timestamps were utilized. Features that were extracted are Kurtosis, Skewness, Entropy, Min, Mean, and Max for Acceleration data, BVP data, and Gyroscope data, amongst others. For EEG data, the main features extracted were power spectral densities, band-powers, band ratios at the different electrodes, spectral entropy, and various statistical features. For the GSR data, the skin conductance level and the skin conductance response value were extracted, amongst others. For the PPG data, the heart rate, heart rate variability, and others were extracted. As a window length of features, an epoch duration of five seconds was utilized. The source code for the data storage and feature extraction was made publicly available at [38].

E. Machine Learning

As a final step, Machine Learning (ML) models were trained to distinguish between the activities performed by the study participants. As ML models, the widely used model-families Logistic Regression (LR), Decision Trees (DT), Nearest Neighbors (NN), and Support Vector Machines (SVM) were employed. Effectively, the ML models were trained as generalized binary activity classifiers. The hyperparameters for each ML model were derived using a nested 5-fold cross-validation scheme, training and evaluating the model performance for a given set of hyperparameters and testing the generalization capabilities on a held-out test set. Hyperparameters for the LR were penalty (*l1, l2, None*) and solver (*lbfgs, liblinear, sag, saga*), for the DT were criterion (*gini, entropy*), splitter (*best, random*, and max_depth (*5, 10, ..., 300, None*)), for the SVM (Linear Support Vector Classifier) were penalty (*l1, l2*), as well as the regularization parameter C (*0.01, 0.1, 1,* *10, 100, 1000*), and for the NN were (leaf_size (*1, 2, ..., 50*), n_neighbors (*1, 2, ..., 30*), and p (*1, 2*)). The train-validate-test split was 60%-20%-20%, and as outer stratified 5-fold CV was employed, while the experimental HalvingGridSearchCV from scikit-learn was utilized for the inner CV [39].

Finally, the resulting performances were averaged and the best hyperparameters were noted down. Due to data imbalances, (*41%*:*59%* for *Cognitive-Load*:*Intervention*), the data was once randomly resampled before the experiments, resulting in balanced data sets.

IV. RESULTS

A. Machine Learning

The mean age of the ten participants was 27.6 years, with a standard deviation of 4.34 years. Due to the sickness of one pair of participants, their respective fifth recording could not be performed, and as such a total of 48 data recordings (24 sessions) were performed. After hyperparameter tuning utilizing nested 5-fold CV and HalvingGridSearchCV, the following hyperparameters were utilized across most of the model runs: for the LR (*penalty = None, solver = lbfgs*), for the DT (*criterion = entropy, splitter = best, max_depth = 145*), for the SVM (*penalty = l1, C = 1000*), and for the NN (*leaf_size = 25, n_neighbors = 21, p = 1*). The resulting model performances for distinguishing between cognitive load induction and yoga intervention are detailed in Table I. The mean across nested CVs (Generalized) or across nested CVs and across participants (Personalized) is reported. The feature sets utilized contained *Kurtosis, HRV, HR, SCL, SCR_Freq*, and *Temp* features (E4), *AF7_alpha_power, AF8_alpha_power, TP9_alpha_power, TP10_alpha_power, AF8_theta_delta, AF7_theta_delta, AF7_low_beta, AF8_low_beta, tfr_9Hz, tfr_18Hz, tfr_27Hz, entropy_AF8, entropy_AF7, entropy_TP9*, and *entropy_TP10* features (Muse), or all of these (All). As can be seen, both for Generalized and Personalized models, the NN (printed in boldface) performed best, while overall DT performed worst. The best performance was consistently achieved using the feature set All, followed by the Muse features, and finally by the E4 features.

An exemplary visualization of some features averaged across all participants is given in Figure 4. Feature values were averaged per participant across the min-max normalized values (and for the EEG features at the electrode positions AF7, AF8, TP9, and TP10), and averaged across recordings. As can be seen, the Alpha Power, which correlates positively with relaxation [40], is increased during the yoga intervention when compared to the cognitive load induction phase, validating its use as a biomarker for cognitive demands. While the heart rate does not seem to change significantly between conditions, the SCL is higher during the intervention than during the load induction at rest. The strong distortion in the physiological signals around the time of the transition from one phase to another, including a lot of uncontrolled movements, is reflected in the data.

Model	Feature-Set	Generalized	Personalized
NN	All	88.80%	**90.01%**
NN	Muse	84.28%	**86.59%**
NN	E4	72.64%	**79.68%**
LR	All	80.12%	**82.13%**
LR	Muse	68.94%	**80.77%**
LR	E4	73.36%	**73.81%**
SVM	All	79.73%	**81.33%**
SVM	Muse	68.40%	**80.45%**
SVM	E4	73.32%	**73.98%**
DT	All	78.06%	**78.60%**
DT	Muse	71.62%	**79.51%**
DT	E4	68.42%	**75.75%**

Table I: Classification accuracy for binary classification models on well-balanced (*50%*:*50%* datasets for *cognitive load*:*yoga*), derived after nested 5-fold Cross-Validations (CV). Generalized models were built using data from all participants while personalized models utilized data from only one participant (no outliers removed). Overall model performances are color-coded from best (blue) to worst (red). The best performance per row is printed in boldface.

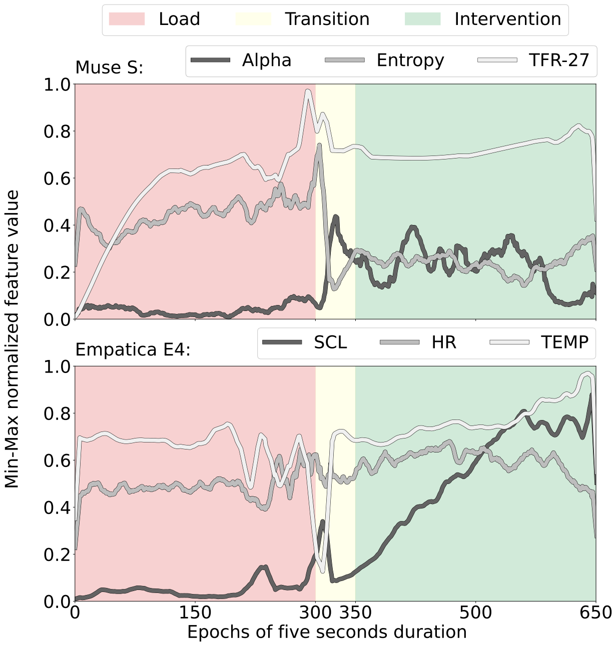

Figure 4: Generalized mean feature values for mean power in the *alpha frequency band (8 - 12 Hz; Alpha), spectral entropy (Entropy)*, and the *spectral power at 27 Hz (TFR-27)* derived for the Muse S (top), and the *skin conductance level (SCL), heart rate (HR)*, and *skin temperature (TEMP)* derived for the Empatica E4 (bottom). The signal was smoothed over twelve consecutive epochs of five seconds, i.e., over one minute.

B. Similarity Analysis

To confirm the validity of synchrony of the recorded EEG data, initially a comparison of the SSVEPs after Oddball paradigm had been planned. However, due to the issues outlined in the Subsection *Similarity*, only two recordings out of the total of 48 data recordings could be considered for the analysis of Event-Related Potentials (ERPs). Figure 5 visualizes these results, which are not generalizable as the analysis was performed only on a few data points. Due to technical difficulties with the oddball presentation paradigm outlined in the Subsection *Similarity*, the data shown is averaged over one session of two participants, respectively. The well-studied ERP components N200 and P300 are well-visible for the oddball paradigm. While the absence of the P300 in the *control* task is expected [13], it is unexpected that no N200 ERP is visible. As a result of the technical difficulties, the absence of the N200 in the *control* task, and the low number of samples, the reliability of ERP analysis on this data is limited. However, in line with related work [13], these results underline the possibility of researching SSVEPs with the utilized low-cost sensors.

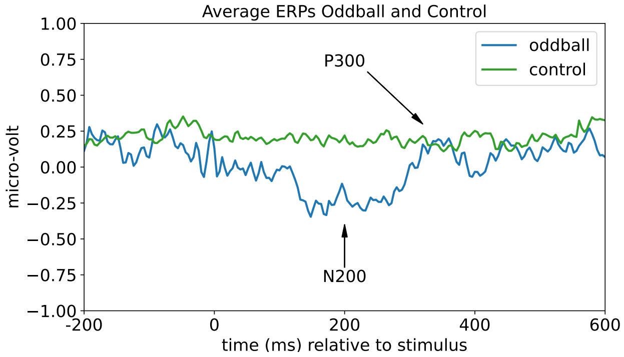

Figure 5: Event-Related Potential (ERP) analysis after oddball paradigm during *oddball* (*blue*) and *control* (*green*) tasks, respectively.

Another approach to analyzing the similarity of the recorded physiological signals is utilizing a distance-based measure for the features extracted from the physiological data [41]. In this work, a Python implementation for Dynamic Time Warping (DTW) was utilized [42]. Compared to some other similarity measures, DTW allows for non-linear matching by stretching or shrinking the compared signals [43], and has also been explored in ML [44]. One challenge for this analysis is that the participants were instructed via video-based yoga to perform the same movements. Consequently, during the recordings, participants did not necessarily perform the same exercise the same way at the same time after an instructed change of pose. A non-linear distance-based measure, such as DTW, is well-suited for this analysis [42]. Here, the normalized DTW distance across the Empatica E4 feature sets *Kurtosis, HRV, HR, SCL, SCR_Freq*, and *Temp* features, and across the Muse S feature sets *AF7_alpha_power, AF8_alpha_power, TP9_alpha_power, TP10_alpha_power, AF8_theta_delta, AF7_theta_delta, AF7_low_beta, AF8_low_beta, tfr_9Hz, tfr_18Hz, tfr_27Hz, entropy_AF8, entropy_AF7, entropy_TP9*, and *entropy_TP10*, was computed between each epoch of each recording. Special interest was placed on enabling the comparison between the pairs of participants. The results are visualized in Figure 6. The color-coded boxes represent the distance within a group of participants, across all sessions. The white box represents the distance of all participants not within the same group, across all sessions. Boxes start at the mean distances during cognitive load and yoga sessions, and their width and height are given by the respective standard deviations. As can be seen, the distances within the groups are smaller than between participants from different groups, but with a high standard deviation. Across participants and groups, the Standard Deviation (STD) of the mean normalized distance across all features and epochs is smaller than the STD over all Muse S features. Generally, the distances within the groups are smaller than the distances between the individuals of the respective group and other recordings.

C. Feature Importance

The importance of individual features was investigated using a correlation analysis performed after artifact removal. To remove the artifacts, a dynamic Interquartile Range (IQR) method built on the STD in each feature was utilized. Details can be found in the source code at [38]. Especially the statistical features extracted from the EEG data (correlation over 0.58 at p-values under 0.001), the heart rate variability (correlation of 0.51 at p-value under 0.001), and the skin conductance level (correlation of 0.45 at p-value under 0.001) were highly correlated with the phase.

D. Limitations

As the technical framework was constantly developed once a bug or a sub-optimal solution was noticed, some recordings produced slightly different artifacts than others. As a result thereof and of issues encountered in the uncontrolled environments, such as a vast amount of Bluetooth devices present in the immediate neighborhood, three recordings show a significant amount of artifacts, and one out of these recordings stored data for all modalities only at a maximum sampling rate of 10 Hz. Generally, due to the nature of the bodily exercise of yoga, the second half of the recordings is partially distorted due to strong movement artifacts when participants changed their yoga poses (only during said change). Furthermore, data labelling during yoga was impractical, as it would have interfered with the participants performing the stress-reducing intervention. Consequently, the temporal resolution of self-assessed labels is significantly higher for the cognitive load task than for the relaxation intervention task. Finally, the recordings were performed in winter, and some participants reported feeling a bit sick. Therefore, some participants asked for the windows to be closed, while other participants appreciated open windows, potentially influencing the comparability of temperature and GSR readings across recordings.

V. FUTURE WORK

Due to the richness of the dataset collected, some aspects remain to be analyzed further. The synchronicity of physiological responses during cognitive load induction, but especially

Copyright (c) IARIA, 2024. ISBN: 978-1-68558-167-1

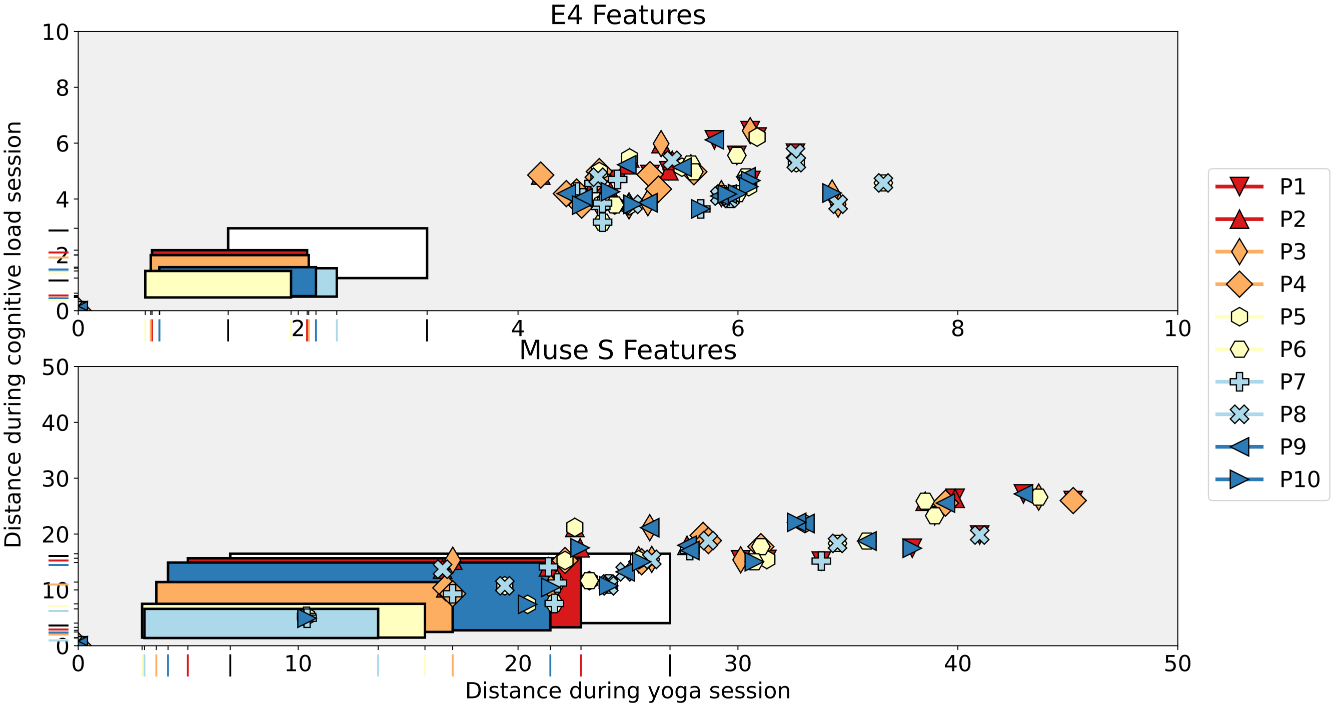

Figure 6: Mean normalized Dynamic Time Warping distances across features from the Empatica E4 (top) and Muse S (bottom), respectively [42]. The pairings of distances are given for each session for the participants not within the same group (i.e., session one of participant P1 was compared to all the 1st sessions of all other participants but the group-partner of P1) and labeled with the markers in the legend.

during the stress reduction mechanisms, should be investigated further. By performing subsequent data collection using the same protocol on individuals rather than on small groups, the stress reduction as determined by the biomarkers could be analyzed and compared, potentially leading to tangible recommendations for organizations' policies. Personalizing the analysis even further, it would be possible to conduct the same analyses and ML regressions using the participant-given labels. If the binary classifiers were trained on the actual user-perceived labels and not on predefined task labels (data available), the classification results are expected to be different. Lastly, ML and Deep Learning models existing in related work could be further personalized on this dataset, and the resulting models could be made publicly available while investigating the effective usefulness of ML and DL compared with traditional statistics.

VI. CONCLUSION

This study's findings on biomarkers of cognitive demands and their ease of use for ML classifiers have significant implications for Personalized eHealth, particularly regarding the development of personalized stress management solutions. Physiological data of five groups of two participants were recorded, following a five-appointment study design. During the appointments, each pair underwent a cognitive load induction and subsequent stress reduction phase. Results show that the sensors are capable of capturing descriptive data. Despite simultaneous task executions, it was found from the similarity analysis that the normalized Dynamic Time Warping distances between extracted features are greater for yoga sessions than during the cognitive load sessions. The derived load classifiers can be integrated into eHealth platforms and offer monitoring or tailored advice on interventions based on the individuals' stress patterns. As such, real-time stress detection would enable immediate suggestions of coping mechanisms like guided breathing exercises or mindfulness meditation prompts. Moreover, the rich dataset of this study, available upon request, offers immense potential for advancing the understanding of stress physiology in real-world applications, which can be leveraged to refine eHealth technologies further, ensuring they meet the unique needs of each individual. This personalized approach not only enhances user engagement but also promises improved health outcomes by addressing stress in a timely and relevant manner and could therefore help shift organizations towards an employee-focused workplace.

ACKNOWLEDGMENT

This research was (partially) funded by the HPI Research School on Data Science and Engineering.

REFERENCES

[1] D. Gopher and E. Donchin, "Workload: An examination of the concept," in *Handbook of perception and human performance, Vol. 2: Cognitive processes and performance*. John Wiley & Sons, pp. 1–49, retrieved: 4, 2024. [Online]. Available: https://psycnet.apa.org/record/1986-98619-019

[2] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (task load index): Results of empirical and theoretical research," in *Advances in Psychology*, ser. Human Mental Workload. North-Holland, vol. 52, pp. 139–183, retrieved: 4, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0166411508623869

[3] S. Delliaux, A. Delaforge, J.-C. Deharo, and G. Chaumet, "Mental workload alters heart rate variability, lowering non-linear dynamics," vol. 10, retrieved: 4, 2024. [Online]. Available: https://www.frontiersin.org/article/10.3389/fphys.2019.00565

[4] A. Holm, K. Lukander, J. Korpela, M. Sallinen, and K. M. I. Müller, "Estimating brain load from the EEG," vol. 9, pp. 639–651, retrieved: 4, 2024. [Online]. Available: http://www.hindawi.com/journals/tswj/2009/973791/abs/

[5] L. Longo, F. Rusconi, and L. Noce, "The importance of human mental workload in web design," in *Proceedings of the 8th International Conference on Web Information Systems and Technologies*. SciTePress - Science and and Technology Publications, pp. 403–409, retrieved: 4, 2024. [Online]. Available: http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0003960204030409

[6] R. L. Charles and J. Nixon, "Measuring mental workload using physiological measures: A systematic review," vol. 74, pp. 221–232, retrieved: 4, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003687018303430

[7] J. M. Morales, J. F. Ruiz-Rabelo, C. Diaz-Piedra, and L. L. Di Stasi, "Detecting mental workload in surgical teams using a wearable single-channel electroencephalographic device," vol. 76, no. 4, pp. 1107–1115, retrieved: 4, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1931720418306597

[8] A. Tonacci *et al.*, "Can machine learning predict stress reduction based on wearable sensors' data following relaxation at workplace? a pilot study," vol. 8, no. 4, p. 448, number: 4 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2227-9717/8/4/448

[9] Y. Pan, X. Cheng, and Y. Hu, "Three heads are better than one: cooperative learning brains wire together when a consensus is reached," vol. 33, no. 4, pp. 1155–1169, retrieved: 4, 2024. [Online]. Available: https://doi.org/10.1093/cercor/bhac127

[10] A. Kuhlen, C. Allefeld, and J.-D. Haynes, "Content-specific coordination of listeners' to speakers' EEG during communication," vol. 6, retrieved: 4, 2024. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnhum.2012.00266

[11] S. Dikker *et al.*, "Brain-to-brain synchrony tracks real-world dynamic group interactions in the classroom," vol. 27, no. 9, pp. 1375–1380, retrieved: 4, 2024. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0960982217304116

[12] D. Bevilacqua *et al.*, "Brain-to-brain synchrony and learning outcomes vary by student–teacher dynamics: Evidence from a real-world classroom electroencephalography study," vol. 31, no. 3, pp. 401–411, retrieved: 4, 2024. [Online]. Available: https://doi.org/10.1162/jocn_a_01274

[13] O. E. Krigolson *et al.*, "Using muse: Rapid mobile assessment of brain performance," vol. 15, retrieved: 4, 2024. [Online]. Available: https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2021.634147

[14] C. Marchand, J. B. De Graaf, and N. Jarrassé, "Measuring mental workload in assistive wearable devices: a review," vol. 18, no. 1, p. 160, retrieved: 4, 2024. [Online]. Available: https://doi.org/10.1186/s12984-021-00953-w

[15] E. Largo-Wight, B. K. O'Hara, and W. W. Chen, "The efficacy of a brief nature sound intervention on muscle tension, pulse rate, and self-reported stress: Nature contact micro-break in an office or waiting room," vol. 10, no. 1, pp. 45–51, retrieved: 4, 2024. [Online]. Available: https://doi.org/10.1177/1937586715619741

[16] C. Finkelstein, A. Brownstein, C. Scott, and Y.-L. Lan, "Anxiety and stress reduction in medical education: an intervention," vol. 41, no. 3, pp. 258–264, retrieved: 4, 2024. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2929.2007.02685.x

[17] S. Sallon, D. Katz-Eisner, H. Yaffe, and T. Bdolah-Abram, "Caring for the caregivers: Results of an extended, five-component stress-reduction intervention for hospital staff," vol. 43, no. 1, pp. 47–60, retrieved: 4, 2024. [Online]. Available: https://doi.org/10.1080/08964289.2015.1053426

[18] M. Sharma and S. E. Rush, "Mindfulness-based stress reduction as a stress management intervention for healthy individuals: A systematic review," vol. 19, no. 4, pp. 271–286, retrieved: 4, 2024. [Online]. Available: https://doi.org/10.1177/2156587214543143

[19] S. A. Smith, "Mindfulness-based stress reduction: An intervention to enhance the effectiveness of nurses' coping with work-related stress," vol. 25, no. 2, pp. 119–130, retrieved: 4, 2024. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/2047-3095.12025

[20] E. Della Valle *et al.*, "Effectiveness of workplace yoga interventions to reduce perceived stress in employees: A systematic review and meta-analysis," vol. 5, no. 2, p. E33, retrieved: 4, 2024. [Online]. Available: https://doi.org/10.3390/jfmk5020033

[21] D. Anheyer, P. Klose, R. Lauche, F. J. Saha, and H. Cramer, "Yoga for treating headaches: a systematic review and meta-analysis," vol. 35, no. 3, pp. 846–854, retrieved: 4, 2024. [Online]. Available: https://doi.org/10.1007/s11606-019-05413-9

[22] M. C. Pascoe, D. R. Thompson, and C. F. Ski, "Yoga, mindfulness-based stress reduction and stress-related physiological measures: A meta-analysis," vol. 86, pp. 152–168, retrieved: 4, 2024. [Online]. Available: https://doi.org/10.1016/j.psyneuen.2017.08.008

[23] D. Hagemann and E. Naumann, "The effects of ocular artifacts on (lateralized) broadband power in the EEG," vol. 112, no. 2, pp. 215–231, retrieved: 4, 2024. [Online]. Available: https://doi.org/10.1016/s1388-2457(00)00541-1

[24] S. L. Kappel, D. Looney, D. P. Mandic, and P. Kidmose, "Physiological artifacts in scalp EEG and ear-EEG," vol. 16, no. 1, p. 103, retrieved: 4, 2024. [Online]. Available: https://doi.org/10.1186/s12938-017-0391-2

[25] K. T. Sweeney, D. J. Leamy, T. E. Ward, and S. McLoone, "Intelligent artifact classification for ambulatory physiological signals," vol. 2010, pp. 6349–6352, retrieved: 4, 2024. [Online]. Available: https://doi.org/10.1109/IEMBS.2010.5627285

[26] D. Gorjan, K. Gramann, K. De Pauw, and U. Marusic, "Removal of movement-induced EEG artifacts: current state of the art and guidelines," vol. 19, no. 1, retrieved: 4, 2024. [Online]. Available: https://doi.org/10.1088/1741-2552/ac542c

[27] W.-J. Lin and H.-P. Ma, "A physiological information extraction method based on wearable PPG sensors with motion artifact removal," in *2016 IEEE International Conference on Communications (ICC)*, pp. 1–6, ISSN: 1938-1883. [Online]. Available: https://ieeexplore.ieee.org/document/7511485

[28] Y. Zhang, M. Haghdan, and K. S. Xu, "Unsupervised motion artifact detection in wrist-measured electrodermal activity data," in *Proceedings of the 2017 ACM International Symposium on Wearable Computers*, ser. ISWC '17. Association for Computing Machinery, pp. 54–57, retrieved: 4, 2024. [Online]. Available: https://doi.org/10.1145/3123021.3123054

[29] B. Alejandro *et al.*, "A comparative study of event-related coupling patterns during an auditory oddball task in schizophrenia," vol. 12, no. 1, p. 016007, retrieved: 4, 2024. [Online]. Available: https://iopscience.iop.org/article/10.1088/1741-2560/12/1/016007

[30] Z. Yan, R. Tan, Y. Li, and J. Huang, "Wearables clock synchronization using skin electric potentials," vol. 18, no. 12, pp. 2984–2998, retrieved: 4, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/8565988

[31] C. Comito, D. Falcone, and A. Forestiero, "Diagnosis detection support based on time series similarity of patients physiological parameters," in *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 1327–1331, ISSN: 2375-0197. [Online]. Available: https://ieeexplore.ieee.org/document/9643300

[32] C. Kothe, B. Venthur, C. Boulay, D. Medine, C. Brunner, and M. Grivich, "Python interface to the lab streaming layer (lsl)," https://github.com/chkothe/PyLSL, accessed May 22, 2024.

[33] S. S. Gadamsetti, "Streamsense," https://github.com/siddhant61/StreamSense, accessed May 22, 2024.

[34] D. Umbricht *et al.*, "Effects of the 5-HT2a agonist psilocybin on mismatch negativity generation and AX-continuous performance task: Implications for the neuropharmacology of cognitive deficits in schizophrenia," vol. 28, no. 1, pp. 170–181, retrieved: 4, 2024. [Online]. Available: https://www.nature.com/articles/1300005

[35] K. O'Keeffe, S. Hodder, and A. Lloyd, "A comparison of methods used for inducing mental fatigue in performance research: individualised, dual-task and short duration cognitive tests are most effective," vol. 63, no. 1, pp. 1–12, retrieved: 4, 2024. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/00140139.2019.1687940

[36] J. Peirce *et al.*, "PsychoPy2: Experiments in behavior made easy," *Behavior Research Methods*, vol. 51, no. 1, pp. 195–203, Feb.

2019, retrieved: 4, 2024. [Online]. Available: https://doi.org/10.3758/s13428-018-01193-y

[37] Kassandra, “30 min beginner yoga,” https://www.youtube.com/watch?v=6hZIzMpHl-c, accessed May 22, 2024.

[38] S. S. Gadamsetti, “Prosense,” https://github.com/siddhant61/ProSense, accessed May 22, 2024.

[39] F. Pedregosa *et al.*, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011, retrieved: 4, 2024. [Online]. Available: https://scikit-learn.org/stable/about.html

[40] K. Mathewson *et al.*, “Regional electroencephalogram (EEG) alpha power and asymmetry in older adults: a study of short-term test–retest reliability,” vol. 7, retrieved: 4, 2024. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnagi.2015.00177

[41] E. Keogh and C. A. Ratanamahatana, “Exact indexing of dynamic time warping,” vol. 7, no. 3, pp. 358–386, retrieved: 4, 2024. [Online]. Available: http://link.springer.com/10.1007/s10115-004-0154-9

[42] T. Giorgino, “Computing and visualizing dynamic time warping alignments in r: The dtw package,” vol. 31, pp. 1–24, retrieved: 4, 2024. [Online]. Available: https://doi.org/10.18637/jss.v031.i07

[43] S. Salvador and P. Chan, “Toward accurate dynamic time warping in linear time and space,” vol. 11, no. 5, pp. 561–580, retrieved: 4, 2024. [Online]. Available: https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/IDA-2007-11508

[44] S. Gudmundsson, T. P. Runarsson, and S. Sigurdsson, “Support vector machines and dynamic time warping for time series,” in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. IEEE, pp. 2772–2776, retrieved: 4, 2024. [Online]. Available: http://ieeexplore.ieee.org/document/4634188/