2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)
October 11-14, 2020. Toronto, Canada

Movement Artifact-Robust Mental Workload Assessment During Physical Activity Using Multi-Sensor Fusion

Abhishek Tiwari[1], Raymundo Cassani[1], Jean-François Gagnon[2], Daniel Lafond[2], Sébastien Tremblay[3] and Tiago H. Falk[1]
[1]*INRS-EMT, Université du Québec, Montréal, Québec, Canada*
[2]*Thales Research and Technology Canada, Québec, Québec, Canada*
[3] *Université Laval, Québec, Québec, Canada*

***Abstract*— Mental workload assessment is of great importance for safety critical applications, especially in situations that involve physical demands, such as with first responders (e.g., paramedics, firefighters, or police officers). Advancements in physiological signal monitoring with wearable sensors have made way for real-time mental workload assessment using physiological signals. However, these models have typically been conducted in controlled laboratory settings and rely on a single physiological modality. As a result, such models often experience a drop in performance due to movement artifacts introduced in real-life conditions. In this paper, we demonstrate that a multi-modal mental workload model not only improves measurement accuracy, but can also increase robustness against physical activity artifacts. To this end, an experiment was conducted where mental workload and physical activity levels were modulated simultaneously while physiological data was collected from 48 participants using off-the-shelf wearable devices. Results show improved mental workload assessment with multi-modal fusion under varying physical activity conditions.**

I. Introduction

Mental workload is a significant variable influencing worker performance [1]. It corresponds to the portion of an individual's mental capacity required by task demands, which can be modified by a variety of performance shaping factors, such as time constraints, environment, or experience [2]. As such, automated mental workload assessment has emerged as a research field to adapt and support, in real-time, complex cognitive work, thus making task execution more efficient.

Mental workload has been traditionally assessed via questionnaires, such as the NASA Task Load Index (NASA-TLX) [3]. However, the use of questionnaires does not allow for continuous mental workload assessment. While increasing the sampling frequency of questionnaires can alleviate some of this limitation, it may compromise user compliance by increasing careless responding, as well as increase user burden [4]. Task-specific quantitative performance measures, in turn, such as reaction time, number of errors, or accuracy have also been used to quantify workload. Still, these metrics require the task to be completed, hence can only be used in a post-hoc manner. More recently, advancements in wearable technologies have made physiological measurement easier and mobile [2], thus allowing for continuous assessment of real-time mental workload with minimal interference to the ambulant user.

Physiological signals reflect the changes of autonomic and central nervous systems and can provide useful information about the users' physical and psychological states. Mental workload has been studied using various physiological measures, such as electrocardiogram (ECG), blood volume pulse (BVP), respiration, galvanic skin response (GSR), blood pressure, ocular measures, and electroencephalogram (EEG). However, most studies have been conducted in controlled laboratory environments with stationary users; the majority has focused on computerized simulations [2].

Physical activity can not only influence the quality of the measured physiological signals [5], but also requires mental resources leading to changes in mental workload [6]. These combined effects make it extremely challenging to measure mental workload in ambulatory settings. For example, intense physical activity effects physiological signals such as decreased heart rate variability (HRV) [7], decreased skin temperature [8] and a shift of electrodermal activity to higher frequency regions [9]. Lastly, physical activity results in artifacts on the signals collected via wearable devices, thus may lead to measurement errors [10]. While such errors may not be crucial for consumer applications, they cannot be disregarded in safety-critical applications.

Recent multi-modal approaches for modelling psychological state have shown to improve prediction performance, as well as provide robustness in real-life conditions involving physical activity [11], [12]. In this paper, we quantify the impact of each modality on overall model performance and show the benefits of a multi-modal system on movement robustness. Experiments are performed with 48 participants performing the NASA Revised Multi-Attribute Task Battery II [13] under two different physical activity conditions and with three different physical workload levels.

978-1-7281-8526-2/20/$31.00 ©2020 Crown

The remainder of this paper is organized as follows: Section II describes the data collection, feature extraction and analysis performed. In Section III, we present and discuss the results obtained. Finally, Section IV presents the conclusions and future directions.

II. Materials and Methods

A. Data Collection

Data was collected from 48 participants (23 female, average age of 27.6 ± 6.6 years old). Participants were instructed to perform tasks from the NASA Revised Multi-Attribute Task Battery II (MATB-II) [13]. Mental workload was modulated at low and high levels by changing the parameters of the MATB-II task. The task consists of three concurrent activities: system monitoring, tracking, and resource management, where the difficulty levels of each of these activities is modulated to control overall mental workload levels.

Alongside the MATB-II task, participants performed physical activity at different levels either on a treadmill or a stationary bike, with 26 participants using the treadmill. Three levels of physical activity were used: rest (no movement), medium (3 km/h: treadmill, 50 rpm: bike), and high (5 km/h, 70 rpm). In total, six combinations of mental and physical workload were tested (2×3). The order of these six sessions was counterbalanced to remove biases introduced by specific sequence of doing physical/mental activity. After each 10-minute session, a 5-minute break was given. During this break, participants were asked to fill the NASA-TLX questionnaire and report their perceived fatigue level on the Borg scale [14]. Overall, the experiment lasted approximately two hours. The experiment protocol was approved by the Ethics Boards of all involved institutions.

Physiological signals were collected during each session. EEG data was collected using a Neurolectrics Enobio 8-channel portable headset [15] at a sampling rate of 500 Hz. Electrode positions according to the 10-20 international system consisted of P3, T9, AF7, FP1, FP2, AF8, T10, and P4; references were placed at Fpz and Nz. Conductive gel was used in the frontal and temporal electrodes, and dry electrodes were employed to collect data from the parietal region. The Bio Harness 3 (BH3) chest-strap recorded ECG (sampling rate 250 Hz) and breathing activity (sampling rate: 18 Hz). Finally, the Empatica E4 wristband was used to collect BVP, skin temperature and GSR. Multi-modal physiological signal collection requires proper synchronization and markers for labelling relevant events; this was managed by the open-source MuLES software [16]. For more details about the data collection and analysis of subjective ratings, the interested reader can refer to [17]. Figure 1 shows the experimental setup in the treadmill condition.

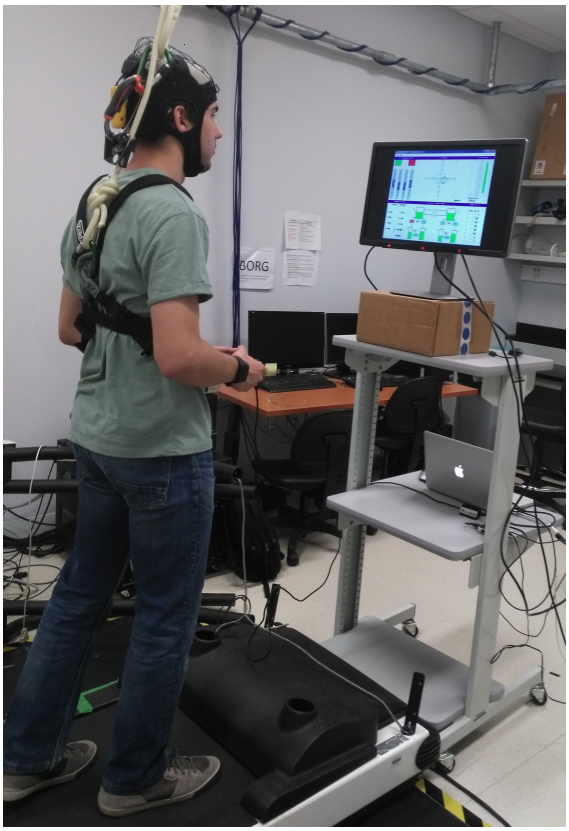

Fig. 1. Experimental setup with the treadmill (standing)

B. Signal Pre-processing

Physiological signals are often corrupted by various artifacts due to motion, other physiological signals, and muscle movements, to name a few [10], [18]. As a result pre-processing artifact removal algorithms are needed. Here, EEG signals were first filtered by a band-pass FIR filter with a bandwidth 1-45 Hz. Following this, ocular and face muscle movement artifacts were filtered using the widely-used wavelet enhanced ICA algorithm [19], [20]. The signal was then decomposed into the following conventional frequency bands: delta (δ, 1-4 Hz), theta (θ, 4-8 Hz), alpha (α, 8-12 Hz), beta (β, 12-30 Hz) and low-gamma (γ_1, 30-45 Hz).

Next, the inter-beat interval (IBI) series was extracted from the ECG signal. First, the ECG was filtered using a 5^{th} order band-pass IIR filter with a bandwidth 4-40 Hz to enhance the QRS complex. This was followed by an energy based QRS detection algorithm [21], which is an adaptation of the Pan-Tompkins algorithm [22]. Visual inspection was performed on a sub-sample of the dataset to ensure beat detection was reliable. The RR series was further filtered to remove outliers using range-based detection ($\geq$ 280 ms and $\leq$ 1500 ms), moving average outlier detection, and a filter based on percent change in consecutive RR values ($\leq$ 20%) as implemented in

[23]. For the breathing signal (BR), downsampling was performed from 18 to 6 Hz followed by filtering the signal using a low pass IIR filter with cutoff frequency of 2 Hz to remove noise.

For the skin temperature signal (TEMP), winsorization [24] (1% - 99% intervals) was first performed for the data to remove any existing outliers in the signal. Winsorization is a statistical method for removing the outlier values over the nth and 100 - nth percentile. This was followed by low pass filtering using a 40^{th} order FIR filter with cutoff frequency of 0.01 Hz to remove high frequency noise. The GSR signal was first down-sampled to 4 Hz. Following this, the phasic high frequency component, associated with sympathetic activity [9], was extracted using a 5^{th} order IIR filter with cutoff frequencies of 0.1-1 Hz. Blood volume pulse was band-pass filtered with a 5^{th} order IIR filter with cutoff frequencies of 8 to 30 Hz to reduce the effects of high frequency noise.

C. Feature extraction

Several benchmark features were extracted for each of the available signals [25], [11], [26]. For EEG, features were computed over 3-second windows (epochs) with 2-second overlap. Spectral sub-band power features were computed and normalized by the full band EEG power. A total of 40 (5 frequency bands $\times$ 8 electrodes) spectral power features were computed per epoch. These features have shown to correlate with mental workload [27].

Next, standard time- and frequency-domain HRV metrics were extracted and used as benchmark ECG measures. Time domain features include mean RR, standard deviation RR (SDRR), coefficient of variation, RMSSD, pNN50, mean of 1^{st} difference, mean of absolute 1^{st} difference and mean of absolute 1^{st} difference of normalized. Frequency domain features, in turn, included low frequency (LF), high frequency (HF), and very low frequency (VLF) powers, normalized LF and HF powers, LF/HF ratio, and total power. The majority of these benchmark features have been shown in the literature to correlate with mental workload [2]. Complete details about these measures can be found in [25]. For respiration signal analysis, statistical descriptors were calculated, namely mean, standard deviation, range, skewness, kurtosis, mean of 1^{st} difference. Additionally, spectral analysis was carried out for the breathing signal by calculating the energy of five equally-spaced bands between 0 to 1 Hz. The spectral energy ratio between 0.05-0.25 Hz and 0.25-0.50 Hz, breathing rate and spectral centroid were also calculated.

For BVP, in turn, the spectrum was divided into five equally-spaced bands between 0 and 2.5 Hz (where much of the signal was contained). Additionally, the spectral energy ratio between 0.04-0.15 Hz and 0.15-0.5 Hz. Skin temperature signal was quantified using statistical descriptors namely, mean, standard deviation, range, mean of 1^{st} difference, min, max, skewness, kurtosis, along with spectral analysis where band energies between 0-0.1 Hz and 0.1-0.2 Hz were also calculated. Finally, for the phasic component of the GSR signal, descriptive statistics were calculated namely, mean, standard deviation, mean of 1^{st} difference, mean of negative 1^{st} difference (MNSSD). Additionally, band power features were calculated for five equally-spaced bands between 0-1 Hz. These bands are investigated as sympathetic activity has been reported to lie in this range with shifts towards higher frequency with physical activity [9]. The features extracted from the different sensors and modalities are summarized in Table I. A total of 94 features were extracted. Unless stated otherwise above, the features were calculated for each session with a 1-minute window and 45-second overlap.

TABLE I
DISTRIBUTION OF FEATURES BY DEVICE AND MODALITY.

Device	Modality	Feature type	Number
Enobio	EEG	Spectral	40
BioHarness 3	IBI	Time- HRV	8
		Frequency - HRV	7
	BR	Descriptive	6
		Spectral	8
E4	GSR	Descriptive	4
		Spectral	5
	BVP	Spectral	6
	TEMP	Descriptive	8
		Spectral	2

D. Classification, and figures-of-merit

Seven feature sets were explored for mental workload evaluation; these feature sets corresponded to the combination of features extracted from various sensors. These included BH3 only, E4 only, and Enobio only; BH3 and E4, BH3 and Enobio, E4 and Enobio combined; and then all devices combined. This was the case for all three physical conditions and explorations for bike and treadmill conditions were evaluated separately. In all cases, the mental workload level (controlled by MATB-II settings) was used as the target labels. A logistic regression model was used for classification. Linear models are simple to interpret and can provide information about feature importance. Additionally, ridge regression has the advantage of being tolerant to high-dimensional datasets. As such, the classification results reported here are to be considered as a lower bound on possible achievable performance with more complex models.

To quantify between-subject generalization, the models were evaluated in a leave-one-subject-out setting. Both epoch- and session-wise evaluations were performed. For epoch-based evaluation, the individual re-

sults for the given epochs for each session were considered as the final result. For session-wise evaluation, a majority voting of the epochs for a given session was made to get a session wise-result. Session-wise evaluation is done as some epochs might be corrupted by sensor noise or distractions of the user. Balanced accuracy (BACC) was used as classifier performance figure-of-merit. The implementation of the classifiers and testing algorithms relied on scikit-learn [28]. Significance of the per-epoch results are gauged by comparing against a random voting classifier.

III. Experimental results and discussion

A. Classification results

The epoch and session-wise performances for bike and treadmill conditions for different physical activity levels are given in Tables II and III, respectively. For the treadmill condition, we see the best epoch-wise performance is achieved with the fusion of all devices for the 'no' and 'high' physical level conditions. The achieved BACC values of 0.704 and 0.645, respectively, show an improvement of 9.3% and 11.2% relative to the the highest performance achieved with an individual sensor (i.e., Enobio and E4). For medium physical level, in turn, the EEG (Enobio) modality showed the highest epoch-wise performance of 0.641, with other sensors showing lower accuracy. When looking at the session-wise majority voting performance, we observe the best no physical activity performance is achieved by two cases with the fusion of all feature sets, as well as by the individual E4 feature set of 0.735. Feature fusion is shown to be extremely important for the medium and high physical activity cases with 0.647 (Enobio + E4, and All) and 0.687 (all) achieved, respectively.

For the bike condition, we see the best epoch wise performance for no and high physical level conditions were from the EEG (Enobio) modality, reaching a BACC of 0.553 and 0.545, respectively. However, for the medium physical activity level, the best performance was achieved with BH3 + E4, with a BACC of 0.617. For the session-wise performance, in turn, we observe best no physical activity performance achieved by three feature sets: BH3, Enobio, and Enobio + E4, all achieving a BACC of 0.547. For medium physical activity, the best performance is achieved by four different feature sets of BH3, E4, (E4 + BH3), and Enobio + BH3 with a BACC of 0.575. Finally, for the high physical activity level, the best performance is achieved from E4 and E4 + BH3 feature sets with a BACC of 0.55.

Overall, the treadmill condition performance was greater than the bike subjects for all three physical activity levels, with differences of 27.3%, 3.9% and 18.3% between best performing models across the no, medium, and high physical activity levels, respectively.

TABLE II

Mental workload prediction performance for different physical activity levels for the treadmill (* represents significance ($p < 0.05$) compared to random voting classifier)

Physical level	Features	BACC (epoch)	BACC (session)
No	BH3	$0.576 \pm 0.086^*$	0.647
	E4	$0.616 \pm 0.118^*$	**0.735**
	Enobio	$0.644 \pm 0.191^*$	0.706
	E4 + BH3	$0.611 \pm 0.126^*$	0.588
	Enobio + BH3	$0.679 \pm 0.186^*$	0.706
	Enobio + E4	$0.662 \pm 0.18^*$	0.647
	All	**0.704 ± 0.188***	**0.735**
Medium	BH3	0.509 ± 0.099	0.470
	E4	0.519 ± 0.11	0.558
	Enobio	**0.641 ± 0.18***	0.617
	E4 + BH3	0.492 ± 0.105	0.529
	Enobio + BH3	$0.573 \pm 0.17^*$	0.617
	Enobio + E4	$0.604 \pm 0.208^*$	**0.647**
	All	$0.568 \pm 0.169^*$	**0.647**
High	BH3	0.553 ± 0.125	0.531
	E4	0.580 ± 0.176	0.625
	Enobio	0.522 ± 0.17	0.531
	E4 + BH3	0.553 ± 0.073	0.562
	Enobio + BH3	0.611 ± 0.159	0.625
	Enobio + E4	0.608 ± 0.18	0.625
	All	**0.645 ± 0.159***	**0.687**

We observe that for most feature sets, session-wise performance was higher than epoch-wise performance for the treadmill condition; they were comparable to each other in case of the bike condition. Session-wise performance is more robust to short duration physiological artefacts as well as subject's distracted mental states which can corrupt epoch-wise results, hence a session-wise output is more useful in a realistic setting. The performance variability in the two conditions could be due to experimental differences between the bike and treadmill conditions. With treadmill activity involving head movement due to walking and running, this could have made focusing on a fixed screen harder compared to more stable upper body position for the bike. This can lead to additional visual processing and attention demands from the user [29], thus making the estimation of mental workload less complex.

Overall, for all per-session conditions for both the treadmill and bike, sensor fusion showed to provide the best BACC, thus showing the importance of fusion for robust mental workload assessment for ambulant users.

B. Feature importance

Analysis of the weights for different features for a logistic regression classifier provides information on how much classifiers relied on individual features. As a result, Tables IV and V show the top-three features from each sensor along with their relative rank in the fused feature set ranking for the no, medium, and high physical activity cases for the treadmill and bike conditions,

TABLE III

MENTAL WORKLOAD PREDICTION PERFORMANCE FOR DIFFERENT PHYSICAL ACTIVITY LEVELS FOR THE BIKE (* REPRESENTS SIGNIFICANCE ($p < 0.05$) COMPARED TO RANDOM VOTING CLASSIFIER)

Physical level	Features	BACC (epoch)	BACC (session)
No	BH3	0.529 ±0.164	**0.547**
	E4	0.486 ±0.158	0.524
	Enobio	**0.553 ± 0.118**	**0.547**
	E4 + BH3	0.486 ±0.186	0.452
	Enobio + BH3	0.523 ±0.142	0.452
	Enobio + E4	0.527 ±0.148	**0.547**
	All	0.500 ±0.185	0.524
Medium	BH3	0.564 ±0.096*	**0.575**
	E4	0.554 ±0.078*	**0.575**
	Enobio	0.471 ±0.151	0.450
	E4 + BH3	**0.617 ± 0.109***	**0.575**
	Enobio + BH3	0.560 ±0.117*	**0.575**
	Enobio + E4	0.486 ±0.149	0.500
	All	0.540 ±0.141	0.500
High	BH3	0.515 ±0.073	0.525
	E4	0.507 ±0.103	**0.550**
	Enobio	**0.545 ± 0.156**	0.500
	E4 + BH3	0.495 ±0.076	**0.550**
	Enobio + BH3	0.524 ±0.142	0.475
	Enobio + E4	0.538 ±0.128	0.525
	All	0.523 ±0.147	0.500

respectively. Feature ranks are defined by the magnitude of the average feature weight across all subjects.

For the treadmill condition, we observe that Enobio (EEG) features are always the highest ranked ones for all physical activity levels. For no physical activity level, the frontal γ energy was shown to be a top feature. Frontal gamma oscillations have been previously linked to engagement in mental activity [30], along with visuo-spatial focused attention [31] and changes during mental arithmetic task [32]. Additionally, θ and δ energies from the right parietal lobe (P4) appeared as top features. This region plays an important role in temporal attention [33]. Functional connectivity between the frontal and parietal regions is related to performance in visual discrimination tasks requiring attention shifts [34].

With increasing physical activity, we see alpha and theta frontal bands become more relevant. An increased theta and decreased alpha power has been reported in high mental workload conditions [27]. Additionally, the most important feature for high physical activity comes from the temporal lobe electrodes. The importance of the temporal lobe electrodes could be due to its role in global visual processing [35], visual discrimination and recognition [36] and also spatial motion and self-motion perception which becomes important with physical activity [37]. For the bike condition, EEG features are among the top features for low and high activity levels. They show similar behavior to the treadmill condition with features from the parietal and frontal electrodes being among the top features. For the medium physical activity case, the EEG features are among the top 6 features with BH3 features being the top features.

Within the BH3 sensor, in turn, top features for the treadmill conditions included SDRR, RMSSD, pNN50, mean of RR, and mean of Δ RR across all conditions. HRV is known to decrease with increasing mental workload, as a result of sympathetic activation and/or to parasympathetic withdrawal [38]. SDRR, RMSSD, and pNN50 are related to the high frequency component of the HRV and hence convey information about parasympathetic withdrawal [25]. With increased physical activity we observe the mean of Δ RR as the top feature. This gives us the mean rate-of-change of the time series and can quantify HF component. The work in [39] reported a significant increase in normalized low frequency power and a decrease in normalized high frequency power with physical activity. However, [40] reported a contradictory increase in the high frequency component with increased exercise intensity on a bicycle. This increase has further been explained by the influence of breathing on heart rate (respiratory sinus arrhythmia, RSA) which has a strong high frequency component during high intensity exercise [41]. Such changes coupled with mental workload could be reflected in the HF component of HRV.

Another top feature observed from BH3 was the standard deviation of BR, which measures the breathing rate variability. Increased sigh rate and respiratory variability have been linked to increased mental workload [42]. Increased sighs are known to induce feelings of relief in subjects and can therefore help perform tasks more efficiently [43]. For BH3 features in the bike condition, in turn, for the no physical activity condition, the pNN50 and mean of normalized absolute Δ of RR series were among the top 20 features, with pNN50 being ranked highly among all physical activity levels. For the medium physical activity, BH3 features appear highly in the ranking, thus corroborating the high performance achieved by the BH3 sensor modality.

Finally, for the E4 sensor, we observe the mean of negative successive differences of GSR signal among the top features across various physical activity levels. This represents the average decrease rate during decay time for the GSR signal [26]. Mental workload levels have been distinguished [44] using GSR fluctuation duration, which is quantified by the decrease rate during decay time. For medium physical activity levels, the top E4 features ranked at the bottom of the list for both the treadmill and bike conditions. Finally, the temperature spectral features are also among the top ranked features for high physical activity levels. Skin temperature changes reflect the sympathetic nervous system activation [45]. Recently, skin temperature has been used to monitor fluctuations of attentions in an arithmetic task [46]. For the bike condition, we see the

top three features are poorly ranked in the overall feature set for no and high physical activity levels. With medium physical activity levels, we see MNSSD (GSR) among the top 20 features which is similar to its performance for treadmill condition.

IV. Conclusion

Physiological models for mental workload assessment have typically relied on single modalities and with stationary participants. Such models, however, do not translate well into real-life settings where subjects are often ambulatory. Here, we show the importance of the fusion of multiple signal modalities to not only improve the performance of mental workload assessment models of ambulant users, but to also provide robustness against movement artifacts. In-depth feature ranking analysis shows the importance of different sensors and signal modalities for the task at hand and may provide future research insights on what sensing modalities to place focus on for mobile workload assessment in highly ecological "in the wild" settings.

For future work, focus is being placed on developing context-aware models with different models for different physical activity levels and types. Additionally, we only made use of a linear logistic regression model; however, more complex models may further improve performance by making use of non-linear decision boundaries. Therefore, other machine learning models (e.g., deep neural networks) should be explored further.

Acknowledgments

The authors acknowledge funding from NSERC, PROMPT Québec, MITACS, and Thales Canada.

References

[1] M. S. Young, K. A. Brookhuis, C. D. Wickens, and P. A. Hancock, "State of science: mental workload in ergonomics," *Ergonomics*, vol. 58, no. 1, pp. 1–17, 2015.

[2] R. L. Charles and J. Nixon, "Measuring mental workload using physiological measures: a systematic review," *Applied ergonomics*, vol. 74, pp. 221–232, 2019.

[3] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research," in *Advances in psychology*. Elsevier, 1988, vol. 52, pp. 139–183.

[4] G. Eisele, H. Vachon, G. Lafit, P. Kuppens, M. Houben, I. Myin-Germeys, and W. Viechtbauer, "The effects of sampling frequency and questionnaire length on perceived burden, compliance, and careless responding in experience sampling data in a student population," 2020.

[5] S. Boettger, C. Puta, V. K. Yeragani, L. Donath, H.-J. Mueller, H. H. Gabriel, and K.-J. Baer, "Heart rate variability, qt variability, and electrodermal activity during exercise," *Medicine & science in sports & exercise*, vol. 42, no. 3, pp. 443–448, 2010.

[6] S. Sharples and T. Megaw, "The definition and measurement of human workload," *Evaluation of human work. Boca*, 2015.

[7] S. Michael, K. S. Graham, and G. M. Davis, "Cardiac autonomic responses during exercise and post-exercise recovery using heart rate variability and systolic time intervals—a review," *Frontiers in physiology*, vol. 8, p. 301, 2017.

[8] G. Tanda, "Skin temperature measurements by infrared thermography during running exercise," *Experimental Thermal and Fluid Science*, vol. 71, pp. 103–113, 2016.

[9] H. F. Posada-Quintero, N. Reljin, C. Mills, I. Mills, J. P. Florian, J. L. VanHeest, and K. H. Chon, "Time-varying analysis of electrodermal activity during exercise," *PloS one*, vol. 13, no. 6, 2018.

[10] D. Tobon, M. Maier, and T. H. Falk, "Ms-qi: A modulation spectrum-based ecg quality index for telehealth applications," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 8, pp. 1613–1622, 2016.

[11] M. Parent, A. Tiwari, I. Albuquerque, J.-F. Gagnon, D. Lafond, S. Tremblay, and T. H. Falk, "A multimodal approach to improve the robustness of physiological stress prediction during physical activity," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. IEEE, 2019, pp. 4131–4136.

[12] A. Tiwari, J. L. Villatte, S. Narayanan, and T. H. Falk, "Prediction of psychological flexibility with multi-scale heart rate variability and breathing features in an "in-the-wild" setting," in *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE, 2019, pp. 297–303.

[13] Y. Santiago-Espada, R. Myer, K. Latorella, and J. Comstock Jr, "The multi-attribute task battery ii (matb-ii) software for human performance and workload research: A user's guide," 2011.

[14] G. Borg, *Borg's perceived exertion and pain scales*. Human kinetics, 1998.

[15] G. Ruffini, S. Dunne, E. Farrés, Í. Cester, P. C. Watts, S. Ravi, P. Silva, C. Grau, L. Fuentemilla, J. Marco-Pallares *et al.*, "ENOBIO dry electrophysiology electrode; first human trial plus wireless electrode system," in *Proceedings of the 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2007, pp. 6689–6693.

[16] R. Cassani, H. Banville, and T. H. Falk, "Mules: An open source EEG acquisition and streaming server for quick and simple prototyping and recording," in *Proceedings of the 20th International Conference on Intelligent User Interfaces Companion*. ACM, 2015, pp. 9–12.

[17] I. Albuquerque, A. Tiwari, J.-F. Gagnon, D. Lafond, M. Parent, S. Tremblay, and T. Falk, "On the analysis of eeg features for mental workload assessment during physical activity," in *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2018, pp. 538–543.

[18] M. K. Islam, A. Rastegarnia, and Z. Yang, "Methods for artifact detection and removal from scalp eeg: A review," *Clinical Neurophysiology*, vol. 46, no. 4-5, pp. 287–305, 2016.

[19] N. P. Castellanos and V. A. Makarov, "Recovering EEG brain signals: Artifact suppression with wavelet enhanced independent component analysis," *Journal of Neuroscience Methods*, vol. 158, no. 2, pp. 300–312, Dec. 2006, 00126.

[20] O. Rosanne, I. Albuquerque, J.-F. Gagnon, S. Tremblay, and T. H. Falk, "Performance comparison of automated eeg enhancement algorithms for mental workload assessment of ambulant users," in *IEEE/EMBS Conf Neural Engineering*, 2019, pp. 61–64.

[21] J. Behar, A. Johnson, G. D. Clifford, and J. Oster, "A comparison of single channel fetal ecg extraction methods," *Annals of biomedical engineering*, vol. 42, no. 6, pp. 1340–1353, 2014.

[22] J. Pan and W. J. Tompkins, "A real-time qrs detection algorithm," *IEEE Trans. Biomed. Eng*, vol. 32, no. 3, pp. 230–236, 1985.

[23] J. A. Behar, A. A. Rosenberg, I. Weiser-Bitoun, O. Shemla, A. Alexandrovich, E. Konyukhov, and Y. Yaniv, "Physiozoo: a novel open access platform for heart rate variability analysis of mammalian electrocardiographic data," *Frontiers in physiology*, vol. 9, p. 1390, 2018.

[24] M. Wu, *Trimmed and winsorized estimators*. Michigan State University, 2006.

[25] A. Camm *et al.*, "Heart rate variability: Standards of measurement, physiological interpretation and clinical use. task force of the european society of cardiology and the north american society of pacing and electrophysiology," *Circulation*, vol. 93, no. 5, pp. 1043–1065, 1996.

TABLE IV

TOP 3 FEATURES FOR EACH SENSOR MODALITY WITH THEIR RELATIVE RANKS FOR EACH PHYSICAL ACTIVITY LEVEL FOR TREADMILL

Sensor	No	Rank	Medium	Rank	High	Rank
Enobio	γ-FP1	1	α-FP2	1	θ-T10	1
	δ-P4	2	β-P4	2	δ-P4	2
	θ-P4	3	α-P4	3	δ-AF7	3
BH3	SDRR (HRV)	11	pNN50 (HRV)	14	std (BR)	9
	RMSSD (HRV)	12	mean RR (HRV)	20	pNN50 (HRV)	10
	std (BR)	13	SDRR (HRV)	28	mean Δ (HRV)	13
E4	std (GSR)	9	energy (1-1.5 Hz) (BVP)	22	MNSSD (GSR)	7
	MNSSD (GSR)	19	MNSSD (GSR)	24	energy (0-0.1 Hz) (TEMP)	15
	energy (1-1.5 Hz) (BVP)	25	mean Δ (TEMP)	35	energy (0.1-0.2 Hz) (TEMP)	16

TABLE V

TOP 3 FEATURES FOR EACH SENSOR MODALITY WITH THEIR RELATIVE RANKS FOR EACH PHYSICAL ACTIVITY LEVEL FOR BIKE

Sensor	No	Rank	Medium	Rank	High	Rank
Enobio	β-FP2	1	γ-P4	2	α-P4	1
	δ-T10	2	β-T9	5	δ-P4	2
	α-P3	3	θ-FP2	6	δ-FP1	3
BH3	pNN50 (HRV)	5	mean δ (HRV)	1	pNN50 (HRV)	12
	mean abs Δ norm (HRV)	20	std (BR)	3	std (BR)	14
	CoV (HRV)	21	pNN50 (HRV)	4	std abs Δ (HRV)	23
E4	energy (0.04-0.06 Hz) (GSR)	28	MNSSD (GSR)	18	std (TEMP)	29
	std (TEMP)	32	energy (0.5-1.0 Hz) (BVP)	38	range (TEMP)	33
	std (GSR)	36	energy (1.5-2.0 Hz) (BVP)	40	energy (1.5-2.0 Hz) (BVP)	37

[26] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physiological signals," *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 18–31, 2011.

[27] G. Borghini, L. Astolfi, G. Vecchiato, D. Mattia, and F. Babiloni, "Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness," *Neuroscience and biobehavioral reviews*, vol. 44, p. 58—75, July 2014. [Online]. Available: https://doi.org/10.1016/j.neubiorev.2012.10.003

[28] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[29] S. Ladouce, D. I. Donaldson, P. A. Dudchenko, and M. Ietswaart, "Mobile eeg identifies the re-allocation of attention during real-world activity," *Scientific reports*, vol. 9, no. 1, pp. 1–10, 2019.

[30] S. Micheloyannis, E. Papanikolaou, E. Bizas, C. J. Stam, and P. G. Simos, "Ongoing electroencephalographic signal study of simple arithmetic using linear and non-linear measures," *International journal of psychophysiology*, vol. 44, no. 3, pp. 231–238, 2002.

[31] J. Kaiser and W. Lutzenberger, "Human gamma-band activity: a window to cognitive processing," *Neuroreport*, vol. 16, no. 3, pp. 207–211, 2005.

[32] R. Ishii, L. Canuet, T. Ishihara, Y. Aoki, S. Ikeda, M. Hata, T. Katsimichas, A. Gunji, H. Takahashi, T. Nakahachi *et al.*, "Frontal midline theta rhythm and gamma power changes during focused attention on mental calculation: an meg beamformer analysis," *Frontiers in human neuroscience*, vol. 8, p. 406, 2014.

[33] S. Agosta, D. Magnago, S. Tyler, E. Grossman, E. Galante, F. Ferraro, N. Mazzini, G. Miceli, and L. Battelli, "The pivotal role of the right parietal lobe in temporal attention," *Journal of cognitive neuroscience*, vol. 29, no. 5, pp. 805–815, 2017.

[34] K. Heinen, E. Feredoes, C. C. Ruff, and J. Driver, "Functional connectivity between prefrontal and parietal cortex drives visuospatial attention shifts," *Neuropsychologia*, vol. 99, pp. 81–91, 2017.

[35] J. Doyon and B. Milner, "Right temporal-lobe contribution to global visual processing," *Neuropsychologia*, vol. 29, no. 5, pp. 343–360, 1991.

[36] R. K. Lech and B. Suchan, "Involvement of the human medial temporal lobe in a visual discrimination task," *Behavioural brain research*, vol. 268, pp. 22–30, 2014.

[37] T. Brandt and M. Dieterich, "The vestibular cortex: its locations, functions, and disorders," *Annals of the New York Academy of Sciences*, vol. 871, no. 1, pp. 293–312, 1999.

[38] G. Chaumet, A. Delaforge, and S. Delliaux, "Mental workload alters heart rate variability lowering non-linear dynamics," *Frontiers in physiology*, vol. 10, p. 565, 2019.

[39] M. P. Tulppo, R. L. Hughson, T. H. Mäkikallio, K. J. Airaksinen, T. Seppänen, and H. V. Huikuri, "Effects of exercise and passive head-up tilt on fractal and complexity properties of heart rate dynamics," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 280, no. 3, pp. H1081–H1087, 2001.

[40] F. Cottin, C. Médigue, P.-M. Leprêtre, Y. Papelier, J.-P. Koralsztein, and V. Billat, "Heart rate variability during exercise performed below and above ventilatory threshold," *Medicine & Science in Sports & Exercise*, vol. 36, no. 4, pp. 594–600, 2004.

[41] G. Blain, O. Meste, and S. Bermon, "Influences of breathing patterns on respiratory sinus arrhythmia in humans during exercise," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 288, no. 2, pp. H887–H895, 2005.

[42] E. Vlemincx, J. Taelman, S. De Peuter, I. Van Diest, and O. Van Den Bergh, "Sigh rate and respiratory variability during mental load and sustained attention," *Psychophysiology*, vol. 48, no. 1, pp. 117–120, 2011.

[43] E. Vlemincx, I. Van Diest, and O. Van den Bergh, "A sigh of relief or a sigh to relieve: The psychological and physiological relief effect of deep breaths," *Physiology & behavior*, vol. 165, pp. 127–135, 2016.

[44] C. Collet, E. Salvia, and C. Petit-Boulanger, "Measuring workload with electrodermal activity during common braking actions," *Ergonomics*, vol. 57, no. 6, pp. 886–896, 2014.

[45] A. Merla and G. L. Romani, "Thermal signatures of emotional arousal: a functional infrared imaging study," in *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2007, pp. 247–249.

[46] T. Lara, E. Molina, J. A. Madrid, and Á. Correa, "Electroencephalographic and skin temperature indices of vigilance and inhibitory control," *Psicológica Journal*, vol. 39, no. 2, pp. 223–260, 2018.