Robust Wearable-based Real Life Cognitive Fatigue Monitoring by Personalized PPG Normalization

Adonis Opris
Mohamed Benouis
Elisabeth André
Yekta Said Can
adonis.opris@uni-a.de
mohamed.benouis@uni-a.de
elisabeth.andre@uni-a.de
yekta.can@uni-a.de
Chair for Human-Centered Artificial Intelligence
University of Augsburg
Augsburg, Germany

Abstract

Cognitive fatigue may have significant results if not intervened in factories, automobiles, and office environments. Development of a system for monitoring cognitive fatigue in real-life settings using unobtrusive wearable devices can help to minimize health problems, and work and car accidents. Photoplethysmography (PPG) sensors offer an unobtrusive way to track changes in heart rate and heart rate variability (HRV), which are indicative of cognitive fatigue levels. In this study, we propose a personalized PPG normalization technique to reduce inter-subject variability and enhance the performance of machine learning algorithms in classifying cognitive fatigue. The best-performing model, a Random Forest Classifier, achieved an accuracy of 80.5% in binary classification and demonstrated robust performance in regression tasks as well. The study highlights the potential of PPG-based wearables for non-obtrusive, long-term monitoring of cognitive fatigue, which could aid in preventing health issues associated with chronic fatigue.

CCS Concepts

• **Human-centered computing → Ubiquitous and mobile computing systems and tools**; **HCI theory, concepts and models**; • **Computing methodologies → Neural networks**; **Supervised learning by classification**.

Keywords

physiological signals, affective computing, cognitive fatigue, machine learning, neural networks, wearable computing.

ACM Reference Format:

Adonis Opris, Mohamed Benouis, Elisabeth André, and Yekta Said Can. 2018. Robust Wearable-based Real Life Cognitive Fatigue Monitoring by Personalized PPG Normalization. In *Proceedings of (IWOAR '24)*. ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

IWOAR '24, June 03–05, 2024, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

When ongoing stress at work isn't handled well, one of the possible outcomes is a cognitive fatigue state. Although it's not a medical disorder, it can happen in any stressful work or home setting and is acknowledged by the World Health Organization (WHO) as a syndrome [15]. In the short term, cognitive fatigue can lead to problems like trouble sleeping, anxiety, irritability, and hormonal issues. Over time, it can cause more serious health problems, including including multiple sclerosis, Parkinson's disease, and traumatic brain injury [6].

Cognitive fatigue can be evaluated by examining the autonomic nervous system (ANS) [18]. One way is to monitor changes in physiological signals. The ANS includes the sympathetic nervous system (SNS) and the parasympathetic nervous system (PSNS). Research shows that physiological signals are good indicators of cognitive fatigue levels over time [10] and [4]. Detecting cognitive fatigue using physiological signals offers several advantages, including non-obtrusive device options, user-friendly systems, objectivity, accuracy, and reliability [3].

Currently, various physiological signals are tested to recognize cognitive fatigue levels. These signals include electrocardiography (ECG), electroencephalography (EEG), photoplethysmography (PPG), electrooculogram (EOG), electromyogram (EMG), electrodermal activity (EDA), skin temperature and respiratory system signals. EEG is one of the most commonly employed modalities for cognitive fatigue because it directly measures the signal from neurons. In one study, researchers collected EEG data obtained from the Muse band during N-back tasks. They obtained 88% accuracy with EEGNet architecture for binary cognitive fatigue detection [5]. ECG signals are also used for cognitive fatigue monitoring and robust performances are obtained. Bhardwaj et al. used ECG with a Stacked Autoencoders algorithm to predict driver fatigue with 90% accuracy based on HRV features [1].

Although robust performances were obtained with ECG and EEG modalities, using them in real life can be inconvenient. It might be plausible to use these devices in short term laboratory experiments but they are not comfortable and obtrusive to use in real life environments such as offices, driving environments, daily life for long term. Users provided chest bands, electrodes and head bands lower comfort, long term use, social acceptance, wearablity and social acceptance scores [16]. At this point, PPG-based devices offer a more practical alternative. They have relatively lower data quality when compared to abovementioned technologies but they provide comfort, social acceptance and aesthetics for long term use. Some studies used mobile phone PPG sensors [18] and wearable PPG sensors[2] for monitoring cognitive fatigue during laboratory experiments.

Another issue with most of the studies in the literature is the usage of tests in laboratory environments. N-back tests, arithmetical tasks, simulations and VR-based tasks are used in laboratory environments. However, these are short-term and artificially applied stimuli. Studies showed that artificially induced affects are different from the ones that occur naturally in the wild [13]. For these reasons, recognizing cognitive fatigue in more realistic environments and tasks for longer terms will contribute more to developing real-life monitoring systems.

In this study, we used a real-life PPG dataset consisting of one-day data [15] from 5 participants and developed a personalized system for recognizing cognitive fatigue levels. We cleaned the PPG signals, extracted distinctive features, employed suitable machine learning algorithms. After that, we employed subjectwise PPG normalization to reduce the effect of interpersonal variability and improved the cognitive fatigue recognition performance. We further provided feature-based analysis and the importance of personalized evaluation in the discussion section.

2 Related Work

As mentioned, the cognitive fatigue monitoring studies using heart activity started in laboratory environments. Lee et al. [8] introduced a method for detecting driver fatigue using 2-minute signal segments from wearable ECG/PPG sensors, achieving a 70% accuracy in binary classification through 10-fold cross-validation. Kundinger et al. [7] developed a non-intrusive fatigue detection system using a wrist-worn ECG sensor, which utilized a 5-minute sliding window with 2-second increments to generate heart rate variability features, achieving a top accuracy of 92.31% for binary classification, comparable to a medical-grade ECG device. Bhardwaj et al. employed the Stacked Autoencoders algorithm to predict driver fatigue based on HRV, attaining a 90% accuracy and noting that HR and LF decreased while HF increased during the transition from alertness to fatigue [1]. Despite the effectiveness of these methods, their reliance on long-term ECG signals hinders real-time performance. To address this, Lei et al. applied a support vector machine (SVM) to short-term ECG signals, each 5 seconds long, for fatigue recognition [9].

After demonstrating the robust performance of ECG-based mental fatigue systems, researchers also tested PPG sensors for the same task. The main reason is the unsuitability of ECG sensors for real life environments especially for long term use. In one study, researchers investigate the real-time detection and prediction of mental fatigue using heart rate (HR) and heart rate variability (HRV) as classification features, extracted from short-term photoplethysmography

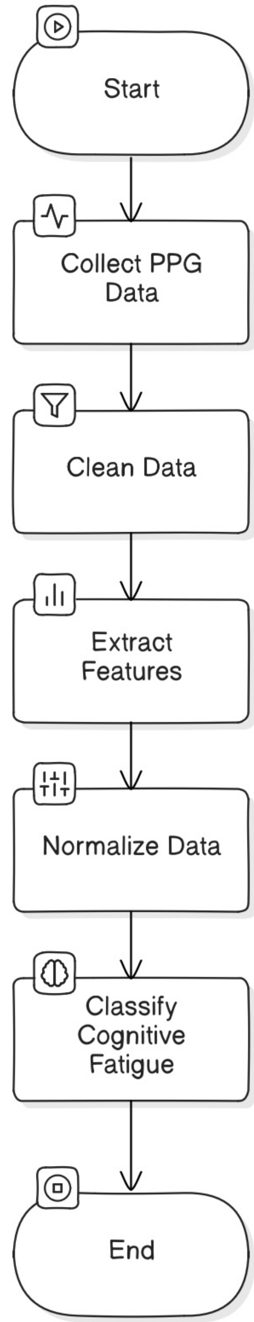

Figure 1: A block diagram of cognitive fatigue monitoring system.

(PPG) signals via smartphones, offering a more convenient alternative to ECG recordings. The researchers processed PPG data by removing baseline wander and smoothing waveforms using polynomial fitting and Savitzky-Golay filtering, then applied an adaptive peak-seeking algorithm to extract R-peaks and calculate HR. Welch spectrum estimation was used to obtain HRV frequency domain characteristics, including high-frequency (HF) and low-frequency (LF) components and the LF/HF ratio. The analysis revealed that HR and HRV features change throughout the day with mental activity, with HR increasing in the afternoon and decreasing in the evening as mental fatigue sets in. The classification of mental fatigue achieved an accuracy of 92.26% and a specificity of 96.12%, indicating that HR and HRV can effectively detect mental fatigue levels in practice, potentially aiding in the prevention of health issues associated with fatigue.

In another study, Alam developed a cognitive fatigue assessment tool, emphasizing the need for contextual evaluation rather than generalized approaches. The authors propose a novel Activity-Aware Recurrent Neural Network (AcRoNN) framework that leverages physiological data from wearable sensors to assess cognitive fatigue. The framework is designed to recognize activities and align them with physiological responses, accounting for artifacts introduced by physical activity. The study evaluates AcRoNN on three datasets, demonstrating significant improvements in cognitive fatigue assessment over baseline models, with a maximum of 19% improvement reported.

Although, there are some recent studies using PPG sensors for real life cognitive fatigue monitoring, the performance of these systems need improvement. In this study, we used the Gamer's Fatigue dataset [15] collected in real-life environments for 24 hours. First, we tested machine learning algorithms for classifying binary cognitive fatigue. We then applied personalized PPG normalization to decrease intersubject differences and showed improvement in terms of performance. We also tested some regression algorithms for levels from 1 to 7. Our results showed the need for alleviating interpersonal variability to obtain more robust performances.

3 Methodology

The analysis process involved the following steps: cleaning the raw PPG signals, dividing them into frames, extracting features that capture the characteristics of the PPG data, labeling the frames based on the self-assessment of sleepiness, and finally, predicting fatigue using machine learning algorithms.

3.1 Preprocessing

The PPG signals were cleaned using the Toolbox for Neurophysiological Signal Processing Neurokit2 [11]. A Butterworth filter was applied with a low-cut frequency of 0.5 Hz [14], a high-cut frequency of 8 Hz [12], and order 3. Figure 2 illustrates the results of cleaning the PPG signal. It can be observed that the noise has been reduced and the baseline corrected.

3.2 Feature extraction

The extraction of features was conducted via a segment-wise analysis, utilising the Heartpy Toolbox [17]. The signal was divided into 300-second segments, with 50% overlap between segments. From each segment, time-domain features were extracted, as well as frequency-domain features. The time domain measures include: beats per minute (bpm), inter-beat interval (IBI), standard deviation

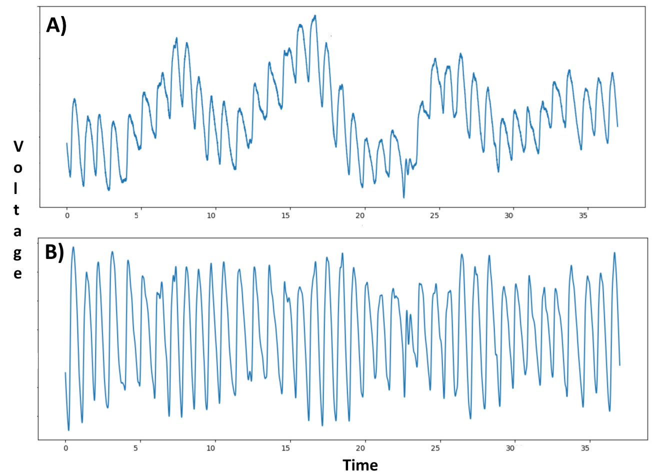

Figure 2: A) PPG signal before cleaning, B) same signal after cleaning.

of NN intervals (SDNN), standard deviation of successive differences (SDSD), root mean square of successive differences (RMSSD), proportion of NN20 (PNN20), proportion of NN50 (PNN50), breathing rate (breathingrate), and some others. The frequency domain features include: very low frequency (VLF), low frequency (LF), high frequency (HF), the ratio of low frequency to high frequency (LF/HF), total power (P_total), percentage of very low frequency (VLF_perc), percentage of low frequency (LF_perc), and percentage of high frequency (HF_perc).

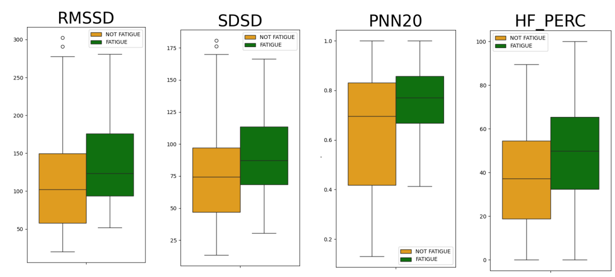

Figure 3: PPG features distribution over the two states, fatigued and non-fatigued

3.3 Feature Level Analysis and Personalized PPG Normalization

Figure 3 shows the distribution of four different features in both fatigue and non-fatigue states. The green box plot represents the fatigue state and includes the frames in which the subject reported the maximum fatigue score. The maximum score varies from subject to subject, there are participants whose maximum fatigue score was 4/7 and others who also used the score 7/7. The yellow ones represents the non-fatiguing state, where participants reported the lowest fatigue score.

When we analyzed the features, the results aligned with the literature [1]. HF component is higher during fatigue state (see Figure 3). These features showed that variance of RR distribution and heart rate variability is higher during fatigue which corresponds to more relaxed conditions as expected.

Once the features had been extracted, the PPG feature values were scaled using the following standardization formula for each subject, where μ is the mean and σ is the standard deviation of the given distribution.

$$z = \frac{x - \mu}{\sigma}$$

In order to train machine learning algorithms, we used the Stanford Sleepiness Scale (SSS) as labels. This was completed by participants after each hour. For regression, we used the scale with its seven values. For classification, we used two classes, with a threshold set at the value of 4 (somewhat foggy, let down) on the SSS scale. This was based on our belief that this is a point where action should be taken to avoid mistakes caused by sleepiness.

3.4 Machine learning models

In order to predict sleepiness, we trained a number of machine learning models with different configurations. For classification, we used a k-Nearest Neighbours (kNN) model with different numbers of neighbours, a Naive Bayes model, an Support vector machine (SVM) model with different kernel functions and C values, a Random Forest (RF) classifier with different depths, and also an MLP model. For regression, we used kNN, Linear Regression (LR), a Random Forest Regressor,and SVM. Table 1 shows the details of the hyperparameters used in the machine learning algorithms.

Table 1: Hyperparameters for Selected Machine Learning Algorithms

Model	Hyperparameter	Values
kNN	N Neighbors	3, 5, 7, 9, 11, 13, 15
SVM	Kernel	linear, poly, rbf, sigmoid
	C	1, 6, 11, 16, 21, 26,31, 36
RF	N Estimators	10, 30, 50, 70, 90
	Max Depth	1, 4, 7, 10, 13, 16, 19, 22, 25, 28

4 Experimental Results

After the pre-processing and feature extraction phase, we removed the rows with missing values and proceeded with training the ML algorithms.

4.1 Classification

The final data set comprised 82.21% instances of fatigue and 17.79% instances of non-fatigue. To ensure the models were not biased towards the majority class, we balanced the data set through random undersampling of the majority class. We trained several machine learning models using the balanced data set and validated them using a five-fold cross-validation process. The best results were achieved with a k-nearest neighbour model, which achieved 61.2% accuracy, and a random forest classifier, which achieved 73.8% accuracy. Figure 5 presents the results obtained with balanced but non-standardised data.

Table 2: Best performing Random Forest Classifier detailed scores

Accuracy	F1 Score	Precision	Recall
80.5	80.5	80.7	80.5

We standardised the data as described in the previous section and trained and validated the models using the same approach with a stratified 5-fold cross-validation. The results of the best-performing machine learning algorithms without balancing the standardised data are shown in Figure 4.

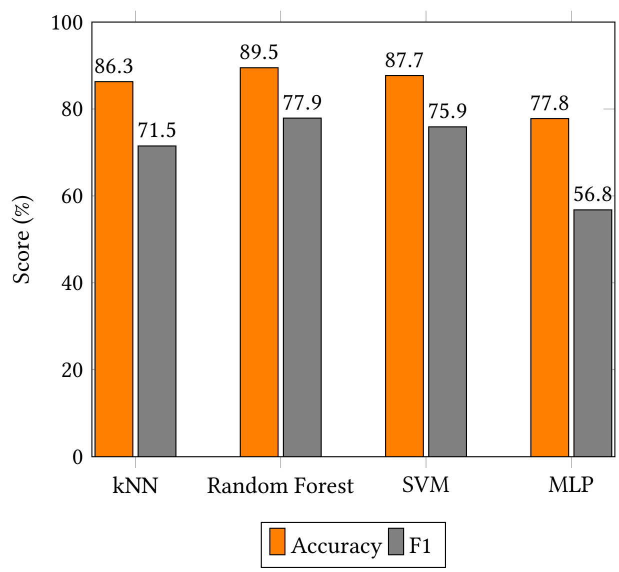

Figure 4: Accuracy and F1 of the best-performing models on unbalanced dataset

After getting results for the original imbalanced dataset, we balanced the dataset by random undersampling and showed the results in Figure 5. As it can be seen, the performance metrics decreased drastically due to losing the advantage of leaning towards majority class.

The results of the best-performing models after balancing and personalized normalization are shown in Figure 6. The kNN achieved 72.6% accuracy with 9 neighbours, the Random Forest Classifier had 80.5% accuracy using 50 trees and a maximum depth of 13, and the SVM used a radial basis function kernel.

The Random Forest Classifier demonstrated the best overall performance in the classification task, which utilized binary classes. Table 2 shows the F1 score, precision, recall for this model, offering additional insights into its performance. Table 3 shows the F1 scores for the Random Forest classifier using different numbers of trees and different depths.

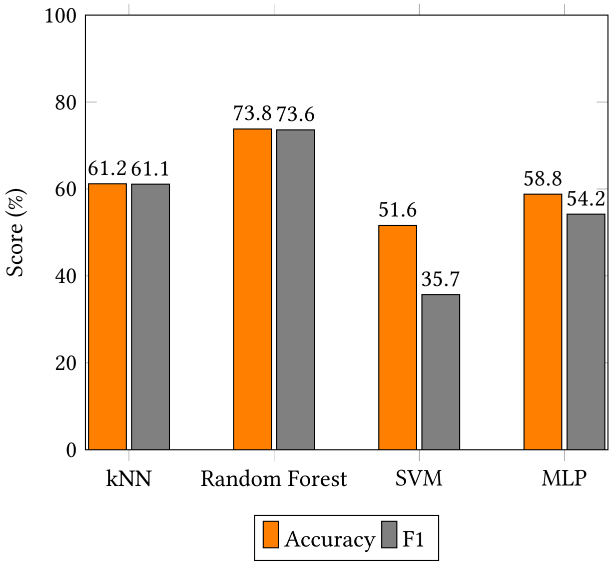

Figure 5: Accuracy and F1 score on non-standardized balanced data set

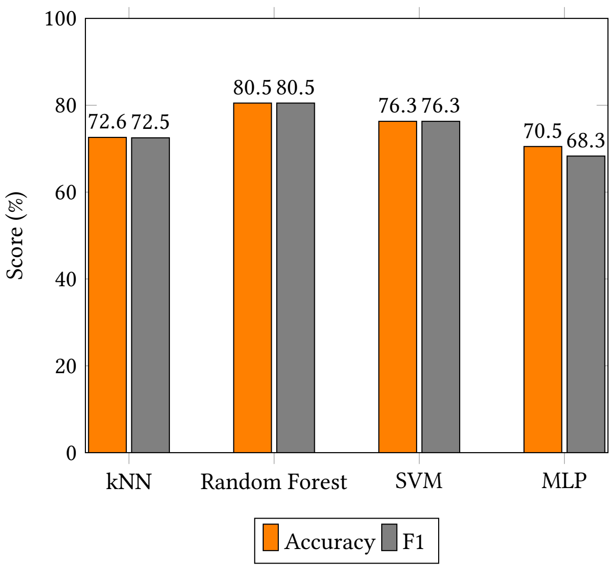

Figure 6: Accuracy and F1 of the best-performing models

4.2 Regression

In order to perform the regression analysis, we utilised the standardised features as input. The target values ranged from one to seven, in accordance with the Stanford Sleepiness Scale. The performance of the models was evaluated using mean squared error (MSE), mean absolute error (MAE), and R2 score, which are presented in Table 4. Similarly to the classification task, in the regression task, the best model was the Random Forest Regressor, this time with 90 trees and a maximum depth of 10.

Table 3: Random Forest F1 scores in percent with different parameter configurations

	Number of Trees				
Max Depth	10	30	50	70	90
1	60.9%	61.3%	63.2%	64.5%	63.4%
4	72.6%	75.7%	76.8%	75.4%	76.5%
7	76.1%	77.7%	77.3%	80.2%	77.2%
10	77.4%	78.4%	78%	79.6%	78.3%
13	77.2%	77.3%	80.5%	78.6%	78.7%
16	76.6%	76.9%	77.4%	77.4%	77.6%
19	75.5%	79.7%	77.1%	78.2%	77.9%
22	76.5%	78.7%	76.7%	77.9%	77.7%
25	75.7%	77.3%	79.1%	78.4%	79.5%
28	77.2%	77.1%	78.3%	78.9%	78.7%

Table 4: Regression models scores

Model	MSE	MAE	R2 score
kNN	1.23	0.84	0.29
Linear Regression	1.49	0.97	0.14
SVM	1.16	0.81	0.33
RF	0.85	0.70	0.51

5 Discussion and Conclusion

We first made a feature level analysis which shows that Parasymphathetic Nervous System activity corresponds to relaxness is increased during fatigue. Our feature analysis confirm this phenomenon with the increase of HF, standard deviation and heart rate variability.

Since our real life dataset is limited in size (24 hours data from 5 participants), we chose traditional machine learning algorithms instead of more complex ones such as CNN, LSTM, Transformers. From the results, it can be seen that Random Forest achieves the best results. We showed both F1 scores and accuracy for both cases. For the imbalanced and general (not personalized) case, accuracy scores are much higher than F1 scores because of ML algorithms' tendency to classify the label as majority class. Therefore, we need to compare the performances before and after personalization from F1 scores. We obtained around 2.6% increase in the binary classification performance for the best result. A different amount of performance increases are demonstrated except for kNN algorithm. We further provided regression results by using the perceived cognitive fatigue scale from 1 to 7. Random Forest Regressor achieves the best results.

We chose a PPG dataset collected with unobtrusive wearables during real world tasks continuously for one day. The reason for that is to obtain more realistic results that can be applied to real world applications. Laboratory induced short term cognitve fatigue will be different than cognitive fatigue caused by a real life task. Furthermore, PPG sensors can be used without disrupting users for longer times. We also improved the performance by applying personalized normalization to decrease the effect of interpersonal variability.

We believe that our system can be applied for especially decreasing cognitive fatigue related work and car accidents by unobtrusive continuous monitoring and giving a chance to intervene before fatigue reaches high levels.

Acknowledgments

This work was carried out within the framework of the AI Production Network Augsburg. This work was partly funded by the German Research Foundation under the project Health-relevant effects of different urban forest structures (LEAF). This work presents and discusses results in the context of the research project ForDigitHealth. The project is part of the Bavarian Research Association on Healthy Use of Digital Technologies and Media (ForDigitHealth), which is funded by the Bavarian Ministry of Science and Arts.

References

[1] Rahul Bhardwaj, Priya Natrajan, and Venkatesh Balasubramanian. 2018. Study to Determine the Effectiveness of Deep Learning Classifiers for ECG Based Driver Fatigue Classification. In *2018 IEEE 13th International Conference on Industrial and Information Systems (ICIIS)*. 98–102. https://doi.org/10.1109/ICIINFS.2018.8721391

[2] Yi-Xiang Chen, Chin-Kun Tseng, Jung-Tsung Kuo, Chien-Jen Wang, Shu-Hung Chao, Lih-Jen Kau, Yuh-Shyan Hwang, and Chun-Ling Lin. 2023. Fatigue Estimation Using Peak Features from PPG Signals. *Mathematics* 11, 16 (2023). https://doi.org/10.3390/math11163580

[3] Carole-Anne Cos, Alexandre Lambert, Aakash Soni, Haifa Jeridi, Coralie Thieulin, and Amine Jaouadi. 2023. Enhancing Mental Fatigue Detection through Physiological Signals and Machine Learning Using Contextual Insights and Efficient Modelling. *Journal of Sensor and Actuator Networks* 12, 6 (2023). https://doi.org/10.3390/jsan12060077

[4] Shitong Huang, Jia Li, Pengzhu Zhang, and Weiqiang Zhang. 2018. Detection of mental fatigue state with wearable ECG devices. *International Journal of Medical Informatics* 119 (2018), 39–46. https://doi.org/10.1016/j.ijmedinf.2018.08.010

[5] Enamul Karim, Hamza Reza Pavel, Ashish Jaiswal, Mohammad Zaki Zadeh, Michail Theofanidis, Glenn Wylie, and Fillia Makedon. 2023. An EEG-based Cognitive Fatigue Detection System. In *Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments* (Corfu, Greece) *(PETRA '23)*. Association for Computing Machinery, New York, NY, USA, 131–136. https://doi.org/10.1145/3594806.3594848

[6] Enamul Karim, Hamza Reza Pavel, Sama Nikanfar, Aref Hebri, Ayon Roy, Harish Ram Nambiappan, Ashish Jaiswal, Glenn R. Wylie, and Fillia Makedon. 2024. Examining the Landscape of Cognitive Fatigue Detection: A Comprehensive Survey. *Technologies* 12, 3 (2024). https://doi.org/10.3390/technologies12030038

[7] Thomas Kundinger, Nikoletta Sofra, and Andreas Riener. 2020. Assessment of the Potential of Wrist-Worn Wearable Sensors for Driver Drowsiness Detection. *Sensors* 20, 4 (2020). https://doi.org/10.3390/s20041029

[8] Jong-Seok Lee, You-Suk Bae, Wongok Lee, Hyunsuk Lee, Jinkeun Yu, and Jong-Pil Choi. 2019. Emotion and Fatigue Monitoring Using Wearable Devices. In *Proceedings of the Sixth International Conference on Green and Human Information Technology*, Seong Oun Hwang, Syh Yuan Tan, and Franklin Bien (Eds.). Springer Singapore, Singapore, 91–96.

[9] Jianmei Lei, Fangli Liu, Qingwen Han, Yunyang Tang, Lingqiu Zeng, Min Chen, Lei Ye, and Li Jin. 2018. Study on Driving Fatigue Evaluation System Based on Short Time Period ECG Signal. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. 2466–2470. https://doi.org/10.1109/ITSC.2018.8569409

[10] Chin-Teng Lin, Mauro Nascimben, Jung-Tai King, and Yu-Kai Wang. 2018. Task-related EEG and HRV entropy factors under different real-world fatigue scenarios. *Neurocomputing* 311 (2018), 24–31. https://doi.org/10.1016/j.neucom.2018.05.043

[11] Dominique Makowski, Tam Pham, Zen J Lau, Jan C Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and SH Annabel Chen. 2021. NeuroKit2: A Python toolbox for neurophysiological signal processing. *Behavior research methods* (2021), 1–8.

[12] Elisa Mejía-Mejía, John Allen, Karthik Budidha, Chadi El-Hajj, Panicos A. Kyriacou, and Peter H. Charlton. 2022. 4 - Photoplethysmography signal processing and synthesis. In *Photoplethysmography*, John Allen and Panicos Kyriacou (Eds.). Academic Press, 69–146. https://doi.org/10.1016/B978-0-12-823374-0.00015-3

[13] Rosalind W. Picard. 2016. Automating the Recognition of Stress and Emotion: From Lab to Real-World Impact. *IEEE MultiMedia* 23, 3 (2016), 3–7. https://doi.org/10.1109/MMUL.2016.38

[14] Kristjan Pilt, Rain Ferenets, Kalju Meigas, Lars-Göran Lindberg, Kristina Temitski, and Margus Viigimaa. 2013. New photoplethysmographic signal analysis algorithm for arterial stiffness estimation. *The scientific world journal* 2013, 1 (2013), 169035.

[15] Mohammad Arif Ul Alam. 2021. Activity-Aware Deep Cognitive Fatigue Assessment using Wearables. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*. 7433–7436. https://doi.org/10.1109/EMBC46164.2021.9630729

[16] Muhammad Umair, Niaz Chalabianloo, Corina Sas, and Cem Ersoy. 2021. HRV and Stress: A Mixed-Methods Approach for Comparison of Wearable Heart Rate Sensors for Biofeedback. *IEEE Access* 9 (2021), 14005–14024. https://doi.org/10.1109/ACCESS.2021.3052131

[17] Paul Van Gent, Haneen Farah, Nicole Van Nes, and Bart Van Arem. 2019. Analysing noisy driver physiology real-time using off-the-shelf sensors: Heart rate analysis software from the taking the fast lane project. *Journal of Open Research Software* 7, 1 (2019), 1–9.

[18] Yaru Yue, Dongjie Liu, Shoudong Fu, and Xiaoguang Zhou. 2021. Heart Rate and Heart Rate Variability as Classification Features for Mental Fatigue Using Short-Term PPG Signals Via Smartphones Instead of ECG Recordings. In *2021 13th International Conference on Communication Software and Networks (ICCSN)*. 370–376. https://doi.org/10.1109/ICCSN52437.2021.9463614

Received 07 July 2024; revised XX July 2024; accepted XX July 2024