Is it a good time to survey you? Cognitive load classification from blood volume pulse

Aneta Lisowska
Institute of Computing Science
Poznań University of Technology
Poznań, Poland
aneta.lisowska@put.poznan.pl

Szymon Wilk
Institute of Computing Science
Poznań University of Technology
Poznań, Poland
szymon.wilk@put.poznan.pl

Mor Peleg
Department of Information Systems
University of Haifa
Haifa, Israel
morpeleg@is.haifa.ac.il

***Abstract*—The CAPABLE project aims to improve the wellbeing of cancer patients managed at home via a mobile Coaching System recommending physical and mental health interventions. Patient reported outcomes are important for evaluation of the efficacy of these interventions. Nevertheless a large number of surveys might be overwhelming to patients. To understand the cognitive demand caused by the surveys and to find the adequate time to prompt patients to complete them we carried out a feasibility study. In this study we developed a machine learning cognitive load detector from blood volume pulse (BVP) captured by a photoplethysmography (PPG) signal. PPG sensors are available on consumer-grade smartwatches, which we will use in our Coaching System. We found that personalised 1D convolutional neural networks trained on raw BVP signal performed better in binary *high vs low cognitive load* classification than the personalised Support Vector Machines trained with heart rate variability and BVP features. We investigated if the further improvements can be obtained by teacher-student semi-supervised model training, nevertheless the performance gains were not notable. In the future we will include additional context information that might aid cognitive load estimation and drive both survey design as well as the timing of the prompts.**

***Index Terms*—Cognitive load, Classification, BVP, mHealth**

I. Introduction

Cancer patients often suffer from a decrease in their physical, mental, and social wellbeing [1]. The Horizon 2020 CAncer PAtient Better Life Experience (CAPABLE) project aims to improve the emotional and physical wellbeing of cancer patients at home. To improve patients' mental wellbeing, the CAPABLE Coaching System delivers via a mobile app evidence-based patient-specific behavioral modification recommendations from the mindfulness, positive psychology, and physical activity domains, which aim to reduce stress, improve sleep, and develop mental resilience. To support the physical wellbeing of patients, the Coaching System delivers clinical guideline-based recommendations that mostly concern treatment of cancer and other comorbidities and monitoring, prevention and treatment of adverse drug events (ADEs). To evaluate the efficacy of the treatments recommended by the Coaching System and to compare them to other treatments, standardized monitoring of patients' quality of life is necessary.

Part of patients' health status can be objectively measured through laboratory test results; other essential parts are complemented by biosignals collected via the patient's smartwatch and through patient-reported outcomes (PROs) collected via the Coaching System's mobile app; the latter are the focus of this paper. PROs [2] are outcomes that are collected directly from patients without an interpretation made by clinicians. They are collected via quantitative survey questions that address different dimensions of wellbeing and allow a standardized way to report the severity of ADEs and other symptoms. For example, National Cancer Institute's Common Terminology Criteria for Adverse Events (CTCAE) [3] includes items for scoring the severity of patients symptoms. The clinical staff might underestimate distress of symptoms experienced by patients [4], therefore PROs are commonly incorporated in clinical studies [5].

The CAPABLE clinical team researched validated PRO instruments and identified ten instruments that the patients should use periodically; each contains 5-30 questions. Among others, they assess outcomes related to nutrition (NRS2002 [6]), sleep patterns (ISI [7]), mood (GAD7 [8], PHQ9 [9]) and quality of life (EORTC QLQ-C30 [10]). Although cancer patients are in general willing to answer questionnaires that could contribute to their own health and to science [11], the large number of such periodic surveys might be overwhelming. We hypothesize that **the appropriate scheduling of the prompts to fill selected questionnaires might impact survey completion**. Patients may be more responsive to a prompt to fill in a survey when they are not occupied by other difficult tasks. Thus prompt scheduling systems should be aware of patients' cognitive load. Inferring cognitive load is commonly performed using electrocardiograms (ECGs) or electroencephalograms (EEGs) [12], however sensors for capturing these signals might be not practical for daily use.

In this feasibility study, we developed a machine learning solution for low cognitive load detection from blood volume

CAPABLE has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 87505

pulse (BVP), which can be captured by consumer-grade smartwatches [13]. The cognitive load detector will guide the timing of the prompts to fill a questionnaire. It could also be used to evaluate if the completion of the questions was too strenuous for the patients.

II. Related Work

A. When are users more receptive to complete surveys?

The number of studies on survey completion rates by cancer patients is limited, therefore we draw on work from other applications. Sarker et al. [14] investigated the relationship between smoking, alcohol use and their mediators. As part of their study, users were prompted to fill in Ecological Momentary Assessment self-reports multiple times daily. They found that the prompts were the least effective when the users were not cognitively available to engage in the activity because they were working or driving a car.

Chan et al. [15] designed a memory coaching app that considers cognitive availability of the user when reminding them to perform a memory exercise. The authors suggested that cognitive availability can be determined by estimation of users' cognitive load and found that users were more receptive to prompts to perform memory exercise under low cognitive load than under high cognitive load.

Based on these results we hypothesise that prompts to fill in a survey delivered when the models detects low cognitive load could be more effective.

B. Cognitive load associated with PRO reporting in CAPABLE

Some of the PRO instruments that the CAPABLE patients will be asked to complete may pose significant cognitive load due to various reasons. First, they ask questions that may raise awareness of difficulties (e.g., "Were you limited in pursuing your hobbies or other leisure time activities?" [10]). Second, some of them contain a large number of questions (e.g., EORTC QLQ-C30 contains 30 questions). Third, standardized symptom reporting by patients, according to the CTCAE terminology, may require patients to perform several steps in order to report a single symptom and its grade. For example first asses the % of area of each body part covered by rash and from this deduce the appropriate symptom severity grade.

The automatic assessment of patients cognitive load during survey completion can guide the stopping criterion on when no more questions should be asked before patient become disengaged.

C. Cognitive load detection from physiological signals

Haapalainen *et al.* [16] collected data from multiple sensors and compared their ability to assess cognitive load. They found that median heat flux measurements in combination with ECG yields most accurate results reaching 80% of accuracy; however each signal on their own had lower performance.

Markova *et al.* [17] trained a Support Vector Machine (SVM) classifier on features extracted from combination of photoplethysmography (PPG) signal with galvanic skin resistance (GSR) and ECG with GSR. The latter yielded better concentration classification performance (78% vs. 74% accuracy). The performance of the model trained on the features from each sensor separately was not reported.

Interestingly, recently a simple measurement of heart rate variability (HRV) calculated as the standard deviation of the time between normal beats (SDNN) and root mean square of successive differences of heartbeat intervals (RMSSD) has been shown to be strongly correlated with self reported cognitive load [18]. Solhjoo *et al.* derived this HRV measures from ECG signal [18], however the HRV can be also captured through measurement of BVP using a PPG sensor [19].

PPG sensors are more commonly found in consumer-grade smartwatches [13] and therefore we intend to develop a model that can detect cognitive load from BVP captured by PPG.

III. Experiment

A. Proposed approach

We are inspired by Chan *et al.* [15] who found that users were more receptive to prompts to perform tasks (memory exercise) under low cognitive load than under high cognitive load. For CAPABLE patients, the tasks are completing PRO questionnaires, which are an essential part of their followup and management. The CAPABLE app will support the entire process: (a) assessing the cognitive availability of patients (i.e., low cognitive load), (b) triggering a pop-up with the questionnaire, and (c) completing the questionnaire. This study focuses on cognitive availability assessment from BVP. The patients will wear smartwatches continuously, yielding a large amount of unlabeled data. On the other hand, the number of labeled examples of high and low cognitive load gathered from each patients will be low, possibly insufficient for supervised training of deep neural network-based models. Therefore, it might be beneficial to leverage from the unlabeled data and adapt semi-supervised approach for cognitive load detector training. In this feasibility study we investigate semi-supervised *teacher-student* approach, which was first introduced by Yalniz *et al.* [20] for image classification task. Here we adapt it to BVP classification (see Figure 1).

The *teacher-student* set up consists of two models and 4 steps:

1) *Teacher model* is trained in a supervised fashion on a small number of labeled examples.
2) *Teacher model* is applied to unlabeled examples. The predictions serve as *pseudolabels*.
3) New model, a *student*, is trained on *pseudolabeled* data
4) *Student model* is fine tuned on the small number of labeled examples.

B. Dataset

We used the publicly available CLAS dataset [17]. The dataset of PPG, EEG and GSR measurements, was gathered from 60 patients involved in cognitively difficult tasks: Stroop test, math test, logic problem test, and emotionally evoking stimuli. The measurements of baseline response were gathered at the start of the session and in-between cognitively or emotionally demanding sessions, when participants were exposed to a neutral stimulus. In this study we consider BVP signals

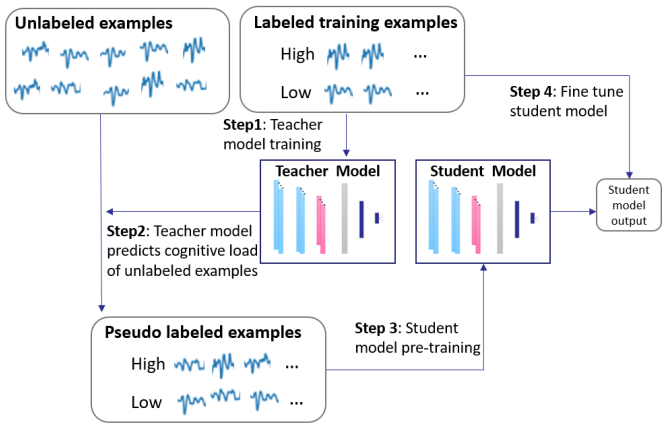

Fig. 1. Teacher-student training procedure.

captured by PPG during high cognitive load tasks (math and logic test) and low cognitive load (baseline and neutral). Note we excluded the Stroop task, which Markova et al. used as a measure of concentration rather than cognitive burden. For each participant cumulatively there are ∼5 minutes of the low cognitive load signal and ∼8 minutes of the high cognitive load signal.

For each subject the unlabeled data used during semi-supervised training consisted of the full signal recordings, this included measurements captured during Stroop test and emotional invoking stimuli. The full length of the recording was ∼35 minutes.

C. Data preparation

Prior to splitting data into training and test datasets we used heartPy [21] Python library to filter the PPG signal. We applied bandpass filter with cutoff frequencies between 0.5 and 3.5 Hz.

We conducted within-patient cross-validation and for each subject we created 3 datasets: labeled training, unlabeled training and labeled testing. We used half of the patient signal from each category (high and low cognitive load) for model training and half for testing. The features were extracted from 10 sec long signal window.

1) Labeled training data: To balance the number of training examples we used varying step size for low and high cognitive load signal. For training the step size for neutral stimuli signal was 5 samples and for high cognitive load 10 samples, yielding $\sim$ 12200 training examples.

2) Unlabeled training data: The step size for unlabeled examples extracted from the full subject recording was 50 samples yielding $\sim$ 12000 unlabeled examples.

3) Test data: For testing the step size was equal to window size for low cognitive load and 1.7 $\times$ window size for high cognitive load to create balanced test set with no overlapping signals windows. This yielded $\sim$ 30 test examples for each subject with $\sim$ 15 examples per class.

D. Machine learning models

1) Support vector machine (SVM): To create baseline performance of the personalised cognitive load classification we trained SVMs using only two features – SDNN and RMSSD, which were previously found to be strongly correlated with cognitive load [18]. Peper *et al.* [19] suggested that there might be more to BVP than the HRV, therefore we also trained SVMs with wider range of statistical features commonly calculated from ECG signal [22]. These are: beats per minute (BPM), the standard deviation of successive differences between neighbouring heart beat intervals (SDSD), SDNN, RMSSD, the proportion of differences between successive heart beats greater than 50ms and 20ms (pNN50, pNN20) and two Poincaré plot measures SD1 and SD2 describing short and long term variability respectively. The features were derived from BVP using heartPy [21]. We used scikit-learn [23] implementation of SVM with default parameters.

2) 1D convolutional neural network (CNN): Both teacher and student model have the same shallow 1D CNN architecture, previously shown effective in stress recognition from BVP signal [24]. We directly input extracted 10-second long snippets of the signal to a simple 1D CNN. The model has two convolutional layers with 16, 8 filters respectively and kernel size of 3, followed by max pooling layer and fully connected layer with 30 nodes and output layer of size 2. Each convolutional layer has a ReLu activation function and output layer with softmax. The model was implemented in Keras with a TensorFlow backend.

E. Evaluation

1) Baseline vs. 1D CNN: Given that the length of training signal for each patients is limited and extracted training examples are heavily over-sampled, we first examined if personalised 1D CNNs could reach baseline performance obtained from personalised SVMs trained with HRV features. Table I shows average accuracy and standard deviation of baseline SVM and 1D CNN across all patients. The performance of the 1D CNN trained on the raw signal examples exceeded performance of the SVM trained with 2 HRV features and with 8 features, suggesting that statistical features do not capture all the information from the BVP signal. Note that the performance of personalised cognitive load classifiers varied between patients, regardless of the classification method.

TABLE I
BASELINE VS. 1D CNN

	SVM (2 features)	SVM (8 features)	1D CNN
Average Accuracy	0.570	0.587	0.605
Std	0.102	0.105	0.075

2) Teacher-student approach: We trained each personalised model 3 times with different random seeds used for model initialisation. In table II we report mean accuracy across all patients and standard deviation (std) between patients for each run as well as the average accuracy between runs. On average

the student model performed just marginally better than the teacher.

TABLE II
TEACHER - STUDENT RESULTS

Run	**Teacher Models Mean Accuracy**	**[Std]**	**Student Models Mean Accuracy**	**[Std]**
1	0.605	0.075	0.611	0.081
2	0.612	0.091	0.614	0.084
3	0.603	0.097	0.603	0.083
Average Accuracy	0.607		0.609	
Std	0.005		0.006	

IV. DISCUSSION

Cognitive load classification from BVP signal is a challenging task. The 1D CNN classifier trained in a personalised fashion on raw BVP signal achieved better results than the personalised SVMs trained with statistical features calculated from BVP. Nevertheless, the results did not reach the cognitive load detection performance previously reported by other researchers when using ECG or a combination of signals from various wearable sensors. Furthermore, in our feasibility study on the CLAS dataset, the semi-supervised *teacher-student* training did not yield notable improvements in cognitive load classification performance. The possible explanation for this result relates to either 1) teacher model performance or 2) the utility of unlabeled set:

1) 1D CNN teacher model might have already achieved the ceiling performance that can be obtained when training on the BVP signal and there is no more extra information that could be extracted from the additional dataset. Alternatively, the teacher model performance could be too low and the provided pseudolabels cannot meaningfully drive training of the student model.
2) The unlabeled dataset did not capture the interesting variability in the cognitive load signal as the majority of the signal that has not been already included in the labeled set came from the participants' exposure to emotional stimuli which might be too different from the cognitive load physiological response. It is also possible that the unlabeled dataset was too small to lead to student model improvements. Yalniz *et al.* [20] used a billion of unlabeled examples.

It is possible that further improvement to the cognitive load classification from BVP might be achieved by different design of the signal annotation (labeling) protocol. Binary labels (high vs. low cognitive load) might suffice to determine whether a patient is occupied by a task and hence should not be interrupted by a prompt to start the survey; however, binary labels do not capture the fine-grained level of cognitive burden caused by a user's activity. This is important for determining when to stop asking the patient to complete additional surveys. We would aim at asking patients to complete two short surveys at one time if they have not "maxed out" on their cognitive load as well as their valence (i.e., degree of positive vs. negative emotion) [25], both of which can be detected from BVP. Patients in such state might experience distress from further requests, leading them to be less compliant to complete surveys in the future.

In future work, we will also explore the incorporation of other information that could be gathered either by consumer grade smartwatches or smartphones that can help with the estimation of a patient's cognitive load, valence, and availability. For example, GPS location might aid prompt scheduling; it could assist in detecting when patients are driving a car and should not be interrupted. As another example, to compare alternative designs of symptom reporting interfaces, based on the degree of cognitive load that they pose, the patients' screen can be captured during survey completion to determine which items of the survey took the longest to complete.

REFERENCES

[1] P. Cardy, J. Corner, J. Evans, N. Jackson, K. Shearn, and L. Sparham, "Worried sick: the emotional impact of cancer," *Worried Sick: the emotional impact of cancer*, 2006.

[2] D. Cella, E. A. Hahn, S. E. Jensen, Z. Butt, C. J. Nowinski, N. Rothrock, and K. N. Lohr, "Patient-reported outcomes in performance measurement," 2015.

[3] A. Trotti, A. D. Colevas, A. Setser, V. Rusch, D. Jaques, V. Budach, C. Langer, B. Murphy, R. Cumberlin, C. N. Coleman *et al.*, "Ctcae v3. 0: development of a comprehensive grading system for the adverse effects of cancer treatment," in *Seminars in radiation oncology*, vol. 13, no. 3. Elsevier, 2003, pp. 176–181.

[4] C. Xiao, R. Polomano, and D. W. Bruner, "Comparison between patient-reported and clinician-observed symptoms in oncology," *Cancer nursing*, vol. 36, no. 6, pp. E1–E16, 2013.

[5] G. Kotronoulas, N. Kearney, R. Maguire, A. Harrow, D. Di Domenico, S. Croy, and S. MacGillivray, "What is the value of the routine use of patient-reported outcome measures toward improvement of patient outcomes, processes of care, and health service outcomes in cancer care? a systematic review of controlled trials," *Journal of clinical oncology*, vol. 32, no. 14, pp. 1480–1510, 2014.

[6] J. Kondrup, H. H. Rasmussen, O. Hamberg, Z. Stanga, A. ad hoc ESPEN Working Group *et al.*, "Nutritional risk screening (nrs 2002): a new method based on an analysis of controlled clinical trials," *Clinical nutrition*, vol. 22, no. 3, pp. 321–336, 2003.

[7] C. H. Bastien, A. Vallières, and C. M. Morin, "Validation of the insomnia severity index as an outcome measure for insomnia research," *Sleep medicine*, vol. 2, no. 4, pp. 297–307, 2001.

[8] N. Williams, "The gad-7 questionnaire," *Occupational Medicine*, vol. 64, no. 3, pp. 224–224, 2014.

[9] K. Kroenke and R. L. Spitzer, "The phq-9: a new depression diagnostic and severity measure," *Psychiatric annals*, vol. 32, no. 9, pp. 509–515, 2002.

[10] N. W. Scott, P. Fayers, N. K. Aaronson, A. Bottomley, A. de Graeff, M. Groenvold, C. Gundy, M. Koller, M. A. Petersen, M. A. Sprangers *et al.*, "Eortc qlq-c30 reference values manual," 2008.

[11] P. G. Kluetz, D. T. Chingos, E. M. Basch, and S. A. Mitchell, "Patient-reported outcomes in cancer clinical trials: measuring symptomatic adverse events with the national cancer institute's patient-reported outcomes version of the common terminology criteria for adverse events (pro-ctcae)," *American Society of Clinical Oncology Educational Book*, vol. 36, pp. 67–73, 2016.

[12] R. Xiong, F. Kong, X. Yang, G. Liu, and W. Wen, "Pattern recognition of cognitive load using eeg and ecg signals," *Sensors*, vol. 20, no. 18, p. 5122, 2020.

[13] S. Saganowski, P. Kazienko, M. Dzieżyc, P. Jakimów, J. Komoszyńska, W. Michalska, A. Dutkowiak, A. Polak, A. Dziadek, and M. Ujma, "Review of consumer wearables in emotion, stress, meditation, sleep, and activity detection and analysis," *arXiv preprint arXiv:2005.00093*, 2020.

[14] H. Sarker, M. Sharmin, A. A. Ali, M. M. Rahman, R. Bari, S. M. Hossain, and S. Kumar, "Assessing the availability of users to engage in just-in-time intervention in the natural environment," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2014, pp. 909–920.

[15] S. W. Chan, S. Sapkota, R. Mathews, H. Zhang, and S. Nanayakkara, "Prompto: Investigating receptivity to prompts based on cognitive load from memory training conversational agent," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 4, pp. 1–23, 2020.

[16] E. Haapalainen, S. Kim, J. F. Forlizzi, and A. K. Dey, "Psycho-physiological measures for assessing cognitive load," in *Proceedings of the 12th ACM international conference on Ubiquitous computing*, 2010, pp. 301–310.

[17] V. Markova, T. Ganchev, and K. Kalinkov, "Clas: A database for cognitive load, affect and stress recognition," in *2019 International Conference on Biomedical Innovations and Applications (BIA)*. IEEE, 2019, pp. 1–4.

[18] S. Solhjoo, M. C. Haigney, E. McBee, J. J. van Merrienboer, L. Schuwirth, A. R. Artino, A. Battista, T. A. Ratcliffe, H. D. Lee, and S. J. Durning, "Heart rate and heart rate variability correlate with clinical reasoning performance and self-reported measures of cognitive load," *Scientific reports*, vol. 9, no. 1, pp. 1–9, 2019.

[19] E. Peper, R. Harvey, I.-M. Lin, H. Tylova, and D. Moss, "Is there more to blood volume pulse than heart rate variability, respiratory sinus arrhythmia, and cardiorespiratory synchrony?" *Biofeedback*, vol. 35, no. 2, 2007.

[20] I. Z. Yalniz, H. Jégou, K. Chen, M. Paluri, and D. Mahajan, "Billion-scale semi-supervised learning for image classification," *arXiv preprint arXiv:1905.00546*, 2019.

[21] P. Van Gent, H. Farah, N. Nes, and B. van Arem, "Heart rate analysis for human factors: Development and validation of an open source toolkit for noisy naturalistic heart rate data," in *Proceedings of the 6th HUMANIST Conference*, 2018, pp. 173–178.

[22] K. Kalinkov, V. Markova, and T. Ganchev, "Front-end processing of physiological signals for the automated detection of high-arousal negative valence conditions," in *2019 X National Conference with International Participation (ELECTRONICA)*. IEEE, 2019, pp. 1–4.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[24] A. Lisowska, S. Wilk, and M. Peleg, "Catching patient's attention at the right time to help them undergo behavioural change: Stress classification experiment from blood volume pulse," in *Proceedings of the 19th International Conference on Artificial Intelligence in Medicine (AIME)*, vol. 12721, 2021.

[25] P. J. Lang, "The emotion probe: studies of motivation and attention." *American psychologist*, vol. 50, no. 5, p. 372, 1995.

The final authenticated version is available at https://doi.org/10.1109/CBMS52027.2021.00061