A Systematic Assessment of Operational Metrics for Modeling Operator Functional State

Jean-François Gagnon[1], Olivier Gagnon[2], Daniel Lafond[1], Mark Parent[3] and Sébastien Tremblay[3]
[1]*Thales Research and Technology, Thales Canada, Québec, Canada*
[2]*Department of Electrical and Computer Engineering, Université Laval, Québec, Canada*
[3]*School of Psychology, Université Laval, Québec, Canada*

Keywords: Operator Functional State, Psychophysiological Modeling, Machine Learning, Data Processing.

Abstract: This paper addresses critical issues and reports key findings with regards to the development of participant-generic operator functional state (OFS) models in the context of cognitive work. Conceptually, this research is concerned with the nature of the relationship between the physiological state of individuals and human performance. Participants were physiologically monitored (cardiac, respiratory, and eye activity) during the execution a set of two cognitive tasks – n-back and visual search – for which there were two levels of difficulty. Levels of difficulty were associated with levels of mental workload. Performance on the tasks was also monitored and linked with OFS. Modeling of the relationship between physiological state and OFS involved systematic manipulation of three parameters: (1) size of smoothing window for performance, (2) performance decrement threshold for labelling functional and sub-functional states, and (3) the mode of classification being either prospective or descriptive. Modeling was performed using two types of classifiers. Results show that (1) models that use bio-behavioral data were capable of classifying performance on new participant data above chance, (2) levels of mental workload were better classified than OFS, (3) size of smoothing window had a significant impact on classifier performance, and (4) size of smoothing window, threshold values, and classifier type had a significant impact on sensitivity and specificity. Implications for the use of OFS models in operational contexts are discussed.

1 INTRODUCTION

The recent development of low cost and mobile devices capable of sensing human bio-behavioral activities has sparked a series of research efforts aiming to use such data for the assessment of operator functional state (OFS) in various contexts. OFS refers to "The multidimensional pattern of human psychophysiological condition that mediates performance in relation to physiological and psychological costs" (Carter et al., 2004). Assessment of OFS has great value, especially in safety critical systems where information about the state of operators could support decision makers or closed loop automated systems (Bracken et al., 2014; Dirican and Göktürk, 2011).

There has been significant progress in the development of models of OFS, but it still faces several challenges before such models are used in the field, especially in safety-critical applications. Indeed, transitioning models to uncontrolled conditions has been identified as an important challenge by many (Durkee et al., 2015; Yin and Zhang, 2014). Specifically, two challenges need to be addressed. The first one concerns the constraints associated with the data, both in terms of quality and availability. The second challenge is at the other end of the spectrum and concerns the formalization of the concept of OFS itself. This paper aims to help address these issues.

1.1 Data

In safety critical contexts, several constraints will impede the use of sensors or degrade the quality of the sampled data. It is not always possible to wear something on the head, or the task may be ambulatory by nature which may introduce motion artefacts in the signals. Consequently, and despite the demonstrated benefits of central nervous system

Gagnon, J-F., Gagnon, O., Lafond, D., Parent, M. and Tremblay, S.
A Systematic Assessment of Operational Metrics for Modeling Operator Functional State.
DOI: 10.5220/0005921600150023
In *Proceedings of the 3rd International Conference on Physiological Computing Systems (PhyCS 2016)*, pages 15-23
ISBN: 978-989-758-197-7

sensors in the modeling of OFS and similar concepts (e.g., Hogervorst et al., 2014; Durantin et al., 2013), some contexts will only allow for the collection of peripheral nervous system measures. Even though it may eventually be possible to remove motion artefacts from the signals using novel techniques (e.g., Tobon et al., 2013), there is a value in investigating the possibility to model OFS using only behavioral and peripheral nervous system sensors. But are they sufficient? Is there really OFS information in the data collected by these sensors? Indeed, the vast majority of studies that model OFS in near real-time rely at least if not only on central nervous system sensors.

Previous work has shown that perceptual attention tasks elicited a small but significant increase in breathing rate (Overbeek et al., 2014). Also, in applied settings, such as office-like situations, breathing rate has shown to be positively associated with stress (Wijsman et al., 2013).

Heart rate variability (HRV) is the (ir)regularity of consecutive heartbeats, and has been widely associated with the balance between sympathetic and parasympathetic systems. Among others, HRV was associated with mental overload in a simulated piloting task (Durantin et al., 2013) and stress in musical performance (Williamon et al., 2013).

Eye-related activity may also provide information associated with OFS. Eye-related attributes should however not require a priori knowledge about the visual scene to facilitate use of the model in new contexts. Such attributes include eye velocity, pattern of saccadic and fixation activity (Régis et al., 2012), blink frequency and blink duration.

Altogether, such attributes have shown to be sensitive to levels of workload, but divergent (Matthews et al., 2015). One potential approach to increase specificity is to combine attributes through machine learning techniques in order to discover multi-modal classification rules.

1.2 OFS Ground Truth

But what exactly is this information about the operator that such models attempt to provide? Gaillard (2003) argued that the goal of OFS assessment "is to detect significant deviations from the optimal bio behavioral state that may indicate an enhanced risk for performance degradation". This conceptualization disentangles performance from OFS by introducing the notion of enhanced risk, which is reasonable since performance greatly depends on contextual factors. Indeed, the concept of performance is arguably further away from the bio-behavioral state than the level of workload or fatigue. It is a multi-determined concept that involves a complex combination of psycho-physiological state, task difficulty, and other contextual factors.

Because of this, one of the most studied components of OFS is mental workload (e.g., Durkee et al., 2013; Eggemeier et al., 1991; Wilson and Russell, 2003), which refers to the portion of operator information processing capacity or resources that is actually required to meet system demands. From a theoretical standpoint, the assessment of mental workload is critical since excessive demand on cognitive resources may result in performance degradation (Nourbakhsh et al., 2013). Still, the relationship between mental workload and performance is not straightforward as other factors come into play, such as level of expertise, fatigue, and motivation. Because the relationship between mental workload and performance is not direct, we might be missing the target. Are the predicted levels of workload really associated with enhanced risk of performance degradation?

This raises the question of how to obtain a valid and reliable ground truth of "enhanced risk for performance degradation". One way to achieve this might be to collect data on standardized fundamental tasks and maximize control of contextual factors. In this context, observed performance degradations should mostly be due to individual as opposed to contextual factors and may therefore be used as a ground truth for OFS. Nevertheless, there are still pending issues we discuss and address here.

1.3 Objectives

The main objectives of this study are (1) to demonstrate the feasibility of modeling OFS without the use of central nervous system sensors and (2) to evaluate various operationalization of OFS ground truth and evaluate how OFS models fare in comparison to models of mental workload. Three specific research questions are addressed.

First, performance and physiology may vary on asynchronous time scales. Indeed, a performance decrement can happen very fast, within seconds, whereas some physiological responses may have a slower onset. For instance, it is usually recommended measuring HRV over five-minute time windows (Mendez et al., 2014). Therefore, is the OFS better conceptualized (and classified) as a punctual or a longer term general state?

Second, it is unclear whether physiological responses cause performance decrements, or if it is the other way around (Brouwer et al., 2015). Are the physiological signals able to *predict* performance decrements (i.e., prospective mode) or are they limited to *describing* an ongoing state? If the performance decrement causes a physiological response, it may be hard to predict decrements in advance. Conversely, if physiological patterns lead to performance decrements, it may be reasonable to have some level of predictability.

Finally, to be reliably detectable, a physiological response must have some level of amplitude. It is unclear however if there is a relationship between the amplitude of the performance decrement and the physiological response. Is there a performance decrement threshold that can be associated with physiological patterns regardless of the task? In other words, what is the magnitude of change necessary in performance to be reflected in physiological response, if any? This paper reports a systematic assessment of these issues.

2 METHOD

This study investigates the aforementioned questions using an experimental design that manipulates task difficulty in order to foster performance decrements. Participants perform the tasks while peripheral bio-behavioral data is collected. Then, a series of models was developed to map the relationship between bio-behavioral data and performance. For each model, a new set of parameters was manipulated for the operationalization of performance. Models are tested on new participants in order to assess cross-subject generalization. This section details the key elements of the method, including data collection, parameter selection, and modeling procedure.

2.1 Experiment

2.1.1 Participants

Seventeen volunteers - 9 males, mean age (sd) = 24.58 (3.74) - participated in the experiment. They were recruited on the university campus and received a financial compensation for their participation. Inclusion criteria were having normal or corrected vision and no known health issues.

2.1.2 Design

The experimental design involved two tasks: visual search and N-Back. Each participant completed eight consecutive experimental sessions separated by five-minute breaks to avoid carry over effects of physiological response. Each task comprised two conditions, easy and hard. This was done to ensure variability in performance data of the participants. These conditions were counterbalanced across participants and played twice. Total duration of the experimental sessions including practice sessions and breaks was approximately 90 minutes. Prior to the experimental sessions, participants were trained on each of the two tasks.

Visual Search. Visual search is a computerized task that requires the participants to identify a target letter among a series of distractors (Figure 1). The task requires visually scanning the screen to search for the target letter. The participants select the target by clicking on the letter. Participants performed 60 trials in each experimental session resulting in 240 trials overall.

Figure 1: Visual search is an attentional task that involves an active visual scan of the environment for a specific target (e.g., unrotated vowel [U]) among distractors (e.g., consonant and rotated vowels).

Task difficulty is manipulated by varying the complexity of the rule of the target letter. In the easy condition, the target is a vowel. In the hard condition, the target is an unrotated vowel. Response times are recorded and represent the measure of performance.

N-Back. The N-Back is a computerized task that requires participants to identify a target letter among a series of distractors presented sequentially in time (one every 2 seconds). The participants are presented a series of letters and must tell whether the actual letter is the same (target) or a different one (distractor) from the N previous letter. Participants must answer with the keyboard (i.e., “M” = same, “Z” = different). Participants performed 60 trials in each experimental session resulting in 240 trials overall. See Figure 2 for a schematic representation of the task.

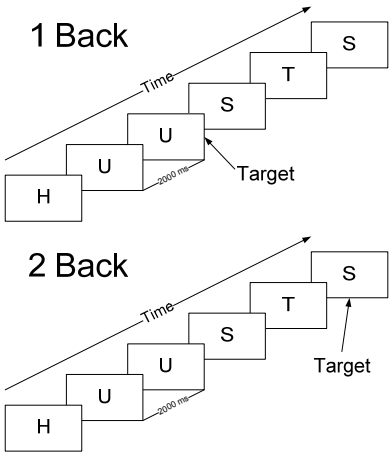

Figure 2: The N-Back task is a working memory task in which the participants must preserve, manipulate and update information in active memory.

Difficulty is operationalized by varying the number of elements to retain, manipulate, and update in memory (i.e., N = 1 [easy] and N = 2 [hard]). Response times are recorded and represent the measure of performance. Accuracy was also recorded; however, we used response times as the principal measure of performance for comparison purposes with the visual search task.

2.1.3 Sensing and Data

During the completion of these tasks, participants were equipped with two devices for bio-behavioral sensing. The Zephyr Bio Harness 3 was used for electrocardiography (ECG) and respiratory induced plethysmography (RIP). ECG was sampled at 250Hz, and RIP was sampled at 18Hz. Data were wirelessly transmitted to the logging device through Bluetooth. The ASL Mobile Eye system was used for eye-tracking. It sampled the position and the dimension of the pupil of the right eye at approximately 30Hz.

Sampled signals were then validated in terms of quality. All data were compared to theoretical boundaries and were investigated further if different from the expected range. Some levels of invalid data were tolerated (up to 20% over a moving window of 10 seconds) to reflect operational conditions. This resulted in approximately 8% of the data removed for model training, validation, and testing.

From the remaining validated signals, a series of attributes were calculated. These attributes are associated with the behavior of the eyes, cardiac activity or respiratory activity.

Eye-related attributes included velocity, proportion of fixations, proportion of involuntary fixations, proportion of saccades, blink frequency, and blink duration.

Several implementations of HRV measures exist and are typically categorized as being either in the temporal, frequency, or non-linear domains (Boonnithi and Phongsuphap, 2011). Since the frequency and temporal domains allow for precise analyses of the variability and was previously associated with mental effort, we adopted these two types. Attributes associated with HRV were extracted using the default values of the RHRV package (Mendez et al., 2014). The frequency bands were extracted over a short window of five minutes and a very short window of two minutes.

In the present study, breathing rate and breathing amplitude were used as the attributes for respiratory activity.

Since the signals were originally sampled at asynchronous rates, we interpolated the values of the attributes with the last valid value between two consecutive samples. The last valid value was used to replicate the functional constraint of the device used for wireless data collection and integration. In fact, all processing is feasible in real-time to acknowledge for operational requirements. All the attributes were then sampled down to 1Hz. The resulting data set was composed of roughly 28,000 observations of 41 attributes.

2.2 Ground Truth Parameters

Decontextualized dynamic performance (DDP) was adapted from previous research and consists of a dynamic standardization of the median response times over the last N seconds (Gagnon et al., 2016). The median response times are standardized so that the resulting score is comparable across tasks. DDP represents the formalization of OFS and allows the direct comparison of multiple tasks in terms of performance.

Aligned with the objectives of the paper, three parameters associated with the calculation (and prediction) of the DDP were manipulated (these parameters are reported in Table 1):

(1) Length in seconds of the median response time window. Two windows were compared: 10 vs. 70 seconds. The shortest window reflects a “punctual” state whereas the longest one represents a “general” state.

(2) Threshold of the z score at which the sub functional level is specified. Three levels were tested: 0, -1 and -1.5. In Gagnon et al. (2016), this parameter was estimated with Yen and colleagues’ method (1995), but its impact was not systematically assessed.

(3) Finally, two types of classification modes were compared: Descriptive of actual state (bio-signals and performance metrics come from the same time window) versus prospective (bio-signals come from a first time window, and performance is assessed using the subsequent time-window).

The manipulation of these parameters results in 12 versions of OFS ground truth which will be systematically assessed in the results section.

Table 1: OFS Parameters.

Sub functional threshold (z score)	Window	
	Punctual (10 sec)	General (70 sec)
0	Descriptive vs. Prospective	Descriptive vs. Prospective
-1	Descriptive vs. Prospective	Descriptive vs. Prospective
-1.5	Descriptive vs. Prospective	Descriptive vs. Prospective

Mental workload, as operationalized by task difficulty (easy vs. hard) was used as an additional ground truth. This ground truth is used to assess the additional difficulty associated with classification of OFS when compared to intermediate psychological states such as mental workload.

2.3 Modeling

In previous work, many classes of models have been used, including support vector machines, decision trees, and linear discriminant analysis, but none have been granted with superior performances (e.g., Gagnon et al., 2016). Because of this, we restrained the classifiers to two types: stochastic gradient boosting machine (GBM) and generalized linear model (GLM). Ensemble methods such as GBM have been used with success in similar contexts (Oh et al., 2015). We compare their performance with GLM, a classic modeling framework that is also known to be resilient to overfitting.

GBM model training was performed using a cross validation procedure implemented in the R caret package (Kuhn et al., 2015). The procedure involved leaving out the data of one participant at a time for training. Data were shuffled prior to input.

The procedure was performed for 4620 (2 X 154 X 3 X 5) iterations: manipulated parameters were interaction depth (2), number of trees (154), shrinkage (3), and minimal observations in node (5).

The threshold used for labelling the data generated imbalanced classes. For instance, the -1 and -1.5 Z score threshold generates sub-functional vs. functional classes comprising ~16% vs. 84% and ~7%vs. 93% of samples, respectively. In order to minimize complications associated with class imbalance, we performed a SMOTE procedure (Torgo, 2010) and classifiers were evaluated using balanced accuracy:

$$Balanced\ Acc. = \frac{(specificity + sensitivity)}{2} \quad (1)$$

This statistic measures classifier accuracy while correcting for class imbalance. It therefore represents a good impartial measure when comparing several scenarios with different class distributions.

For each ground truth (i.e., combination of OFS time window, sub-functional threshold, and mode, and mental workload levels), the best set of parameters was used to train a final model with the data of the 15 training participants. Results report statistics on the testing sample. The test set was divided into eight data bins for each of which balanced accuracy values, sensitivity, and specificity were calculated.

3 RESULTS

Before the development of the psychophysiological models, we first validated that the experimental conditions did have a significant impact on performance of the participants. Analyses revealed that responses times were statistically different between low and high workload conditions for both N-Back $t(16) = 8.29$, $p < .001$ and visual search $t(16) = 10.63$, $p < .001$. The distributions are visually represented in Figure 3 and Figure 4.

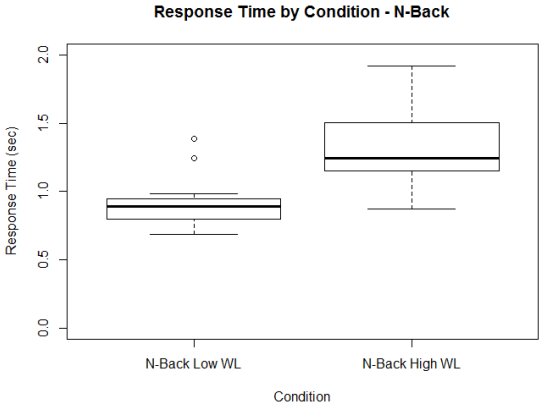

Figure 3: Distribution of response times for the N-Back task by condition.

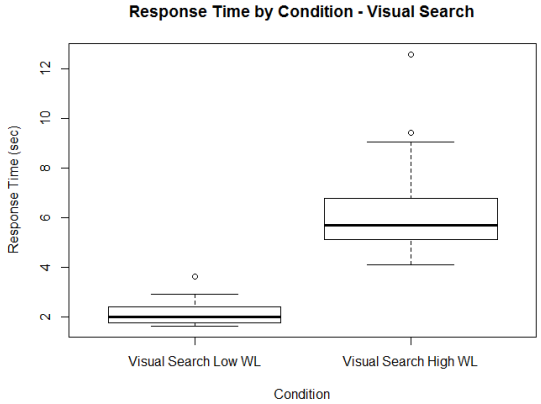

Figure 4: Distribution of response times for the Visual Search task by condition.

Results showed that balanced accuracy was significantly higher for the best mental workload classifier compared to the best OFS classifier $t(15)= 4.85$, $p < .001$. Indeed, the best average balanced accuracy values on the test set was of .77 (.05) for mental workload whereas it was of .66 (.04) for OFS. Moreover, for the best combination of parameters, both mental workload $t(7) = 14.58$, $p < .001$ and OFS $t(7) = 11.29$, $p < .001$ predictive accuracy was significantly superior to chance.

A repeated measures analysis of variance (ANOVA) was carried out to test the effect of classifiers (GBM vs. GLM), time frame (punctual vs. general), mode (descriptive vs. prospective), and threshold (-1.5, -1, 0) on balanced accuracy, specificity, and sensitivity. Statistics are reported in Table 2, 3 and 4 for balanced accuracy, sensitivity and specificity respectively. Data for sensitivity and specificity are reported in Figure 5 and Figure 6 respectively.

Table 2: ANOVA - Effect of threshold, window, mode, and classifier on balanced accuracy.

	df	*F*	*p*
Threshold	2	1.1745	0.3112
Window	1	27.5743	< .001***
Mode	1	3.8706	0.0506
Classifier	1	0.2908	0.5903
Residuals	186		

Table 3: ANOVA - Effect of threshold, window, mode, and classifier on sensitivity.

	df	*F*	*p*
Threshold	2	3.9429	0.0210*
Window	1	3.2754	0.0719
Mode	1	3.349	0.0688
Classifier	1	5.9328	0.0158*
Residuals	186		

Table 4: ANOVA - Effect of threshold, window, mode, and classifier on specificity.

	df	*F*	*p*
Threshold	2	5.2295	0.0061**
Window	1	7.2587	0.0077**
Mode	1	0.4965	0.4819
Classifier	1	17.4499	< .001***
Residuals	186		

Balanced accuracy was not statistically different across OFS classifiers. However, classifiers had an effect on both sensitivity and specificity. Indeed, mean sensitivity was higher for GLM than GBM, and conversely for specificity.

The effect of time window on balanced accuracy was statistically significant. Indeed, observed balanced accuracy on test set was higher for punctual window (M = .59, SD = .07) than general window (M = .54, SD = .07). The time window also had an effect on specificity, but not sensitivity. Indeed, specificity was lower when OFS was conceptualized with a general window than when compared to a punctual window.

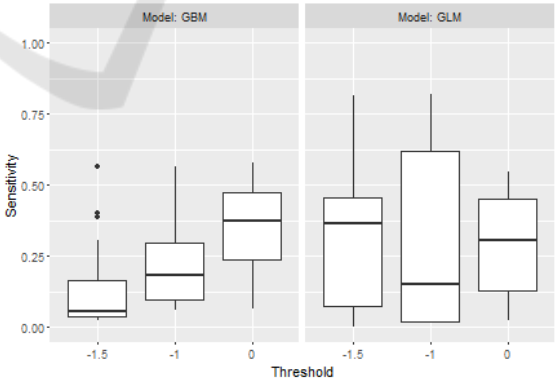

Figure 5: Sensitivity by classifier (left: GBM, right: GLM) and threshold on test data.

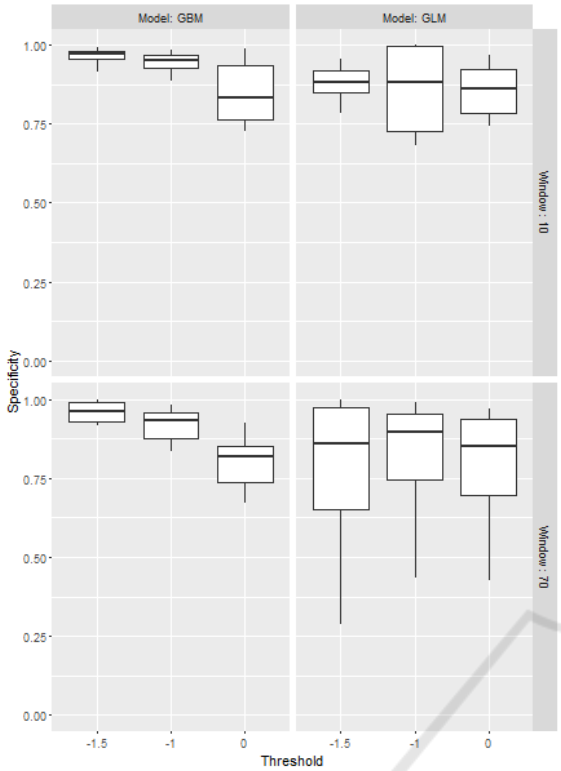

Figure 6: Specificity by classifier (left: GBM, right: GLM), threshold, and time window (top: 10s., bottom: 70 s.) on test data.

There was no effect of threshold on balanced accuracy, but its impact was significant on both sensitivity and specificity. Lower thresholds were on average associated with higher specificity, but lower sensitivity.

Finally, the effect of the classification mode (descriptive vs. prospective) on balanced accuracy was marginally significant. Indeed, balanced accuracy in descriptive mode (M = .58, SD = .07) was almost higher than in the prospective mode (M = .56, SD = .07). There was no effect on specificity and sensitivity.

4 DISCUSSION

This study investigated the potential link between bio-behavioral state of individuals and the likelihood of performance decrements (i.e., OFS). First, we compared model accuracy when classifying performance vs. mental workload. Second, we assessed the impact of key parameters that must be selected when operationalizing an OFS ground truth metric to support supervised learning. Namely, this study investigated if OFS is better conceptualized as a punctual or a general state, as a small or a large variation of performance, and whether a prospective approach can lead to results comparable to the descriptive mode.

As anticipated, results show that the link between the bio-behavioral state of individuals and performance is less strong than the link between that state and intermediate variables such as level of mental workload. Indeed, classifier performance is higher when discriminating workload level compared to OFS level (functional vs. sub-functional). This is not surprising as it was hypothesized before that the concept of performance is dependent on contextual factors not captured in bio-behavioral signals. Moreover, because results show that it is possible to classify workload relatively well, we can affirm that the collected data does in fact carry valid information about the state of the operator.

Results show that it is possible to classify OFS above chance level but that more information is needed to achieve high levels of accuracy. This suggests that if such models are to be used in operational contexts, they should be complemented with other sensor data or contextual information to increase their accuracy. Notably, in this context, we did not provide as inputs the time of the day nor the gender or the age of the participants, which are all known to have an impact on physiological patterns (e.g., Carter et al., 2004). Indeed, without a larger sample and sufficient counterbalancing of these factors, such information will lead to model that overfit the data (e.g., learn the experimental design and idiosyncrasies in the dataset) and that generalize poorly on new data. Still, OFS classifiers with moderate accuracy can be valuable assets since they provide unique information about the individual's state that could help prevent major errors. On the other hand, results also indicate that more work needs to be carried out to boost model generalization. Specifically, while out-of-sample generalization was the priority here, a limiting factor of the current work is that it did not address cross-task generalization even though it is known to be a challenging issue (Wang et al., 2012).

From a theoretical standpoint, results show that OFS is better conceived as a punctual than a general state. Indeed, the short time window has shown to be more easily classified than the long window. This replicates a previous finding observed on a different data set (Gagnon et al., 2016) and suggests that the bio-behavioral signals used for this study (which were all peripheral sensors) are capturing

physiological dynamics that operate on relatively short time scales. Further work should focus on quantifying the prominence of different attributes and their origin (i.e., cardiac, eye, respiratory) with respect to the time window used.

Interestingly, there was no effect of threshold on balanced accuracy. However this parameter, as well as time window and classifier type, had significant impacts on specificity and sensitivity. This is very important, especially in the context of safety critical systems, since you may want to boost one of these metrics over the other. As such, the present findings provide useful insights about parameter tradeoffs and how to prioritize true-positives or true-negatives without compromising balanced accuracy.

Future work will concern training and validation of OFS models in ambulatory contexts, another key challenge to address in order to transit models from the laboratory to the field.

ACKNOWLEDGEMENTS

This research was supported by a Mitacs internship awarded to Mark Parent, funded by NSERC and Thales Canada. The authors would also like to thank Margot Beugniot for her participation in data collection.

REFERENCES

Boonnithi, S., Phongsuphap, S., 2011. Comparison of heart rate variability measures for mental stress detection, *in: Computing in Cardiology, 2011*. pp. 85–88.

Bracken, B.K., Palmon, N., Romero, V., Pfautz, J., Cooke, N.J., 2014. A Prototype Toolkit for Sensing and Modeling Individual and Team State. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting 58*, 949–953. doi:10.1177/1541931214581199.

Brouwer, A.-M., Zander, T.O., van Erp, J.B.F., Korteling, J.E., Bronkhorst, A.W., 2015. Using neurophysiological signals that reflect cognitive or affective state: six recommendations to avoid common pitfalls. *Frontiers in Neuroscience 9*. doi:10.3389/fnins.2015.00136.

Carter, R., Cheuvront, S. N., and Sawka, M. N., 2004. Operator Functional State Assessment (l'évaluation de l'aptitude opérationnelle de l'opérateur humain). Army research institute.

Dirican, A.C., Göktürk, M., 2011. Psychophysiological measures of human cognitive states applied in human computer interaction. *Procedia Computer Science, World Conference on Information Technology 3*, 1361–1367. doi:10.1016/j.procs.2011.01.016.

Durantin, G., Gagnon, J.-F., Tremblay, S., Dehais, F., 2014. Using near infrared spectroscopy and heart rate variability to detect mental overload. *Behavioural Brain Research 259*, 16–23. doi:10.1016/j.bbr.2013.10.042.

Durkee, K.T., Pappada, S.M., Ortiz, A.E., Feeney, J.J., Galster, S.M., 2015. System Decision Framework for Augmenting Human Performance Using Real-Time Workload Classifiers. *Presented at the 2015 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*, Orlando, FL.

Eggemeier, F.T., Wilson, G.F., Kramer, A.F., Damos, D.L., 1991. Workload assessment in multi-task environments. *Multiple-task performance* 207–216.

Gagnon, O., Lafond, D., Gagnon, J-F., and Parizeau, M., 2016. Comparing Methods for Assessing Operator Functional State. *Proceedings of the 2016 IEEE International Inter-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*, San Diego, CA, USA, March 21-25.

Gaillard, A.W., 2003. Fatigue assessment and performance protection, *NATO Science Series Sub Series I Life and Behavioural Sciences 355*, 24–35,.

Hogervorst, M.A., Brouwer, A.-M., van Erp, J.B.F., 2014. Combining and comparing EEG, peripheral physiology and eye-related measures for the assessment of mental workload. *Frontiers in Neuroscience 8*. doi:10.3389/fnins.2014.00322.

Kuhn, M., 2015. caret: Classification and regression training. *Astrophysics Source Code Library 1*, 05003.

Matthews, G., Reinerman-Jones, L.E., Barber, D.J., Abich, J., 2015. The Psychometrics of Mental Workload Multiple Measures Are Sensitive but Divergent. *Human Factors: The Journal of the Human Factors and Ergonomics Society 57*, 125–143. doi:10.1177/0018720814539505.

Nourbakhsh, N., Wang, Y., and Chen, F., 2013. GSR and blink features for cognitive load classification. *In Human-Computer Interaction–INTERACT,* pp. 159-166. Springer Berlin Heidelberg.

Oh, H., Hatfield, B.D., Jaquess, K.J., Lo, L.-C., Tan, Y.Y., Prevost, M.C., Mohler, J.M., Postlethwaite, H., Rietschel, J.C., Miller, M.W., Blanco, J.A., Chen, S., Gentili, R.J., 2015. A Composite Cognitive Workload Assessment System in Pilots Under Various Task Demands Using Ensemble Learning, in: Schmorrow, D.D., Fidopiastis, C.M. (Eds.), *Foundations of Augmented Cognition, Lecture Notes in Computer Science*. Springer International Publishing, pp. 91–100.

Overbeek, T. J., van Boxtel, A., and Westerink, J. H., 2014. Respiratory sinus arrhythmia responses to cognitive tasks: Effects of task factors and RSA indices. *Biological psychology 99*, 1-14.

Régis, N., Dehais, F., Rachelson, E., Thooris, C., Pizziol, S., Causse, M., and Tessier, C., 2014. Formal

Detection of Attentional Tunneling in Human Operator–Automation Interactions. *IEEE Transactions on Human-Machine Systems 44(3),* 326-336.

Rodríguez-Liñares, L., Vila, X., Mendez, A., Lado, M., and Olivieri, D., 2008. RHRV: An R-based software package for heart rate variability analysis of ECG recordings. *In 3rd Iberian Conference in Systems and Information Technologies (CISTI 2008)*, Vigo, Spain.

Tobon D.V., Falk, T., and Maier, M., 2014. MS-QI: A Modulation Spectrum-Based ECG Quality Index for Telehealth Applications. *IEEE Transactions on Biomedical Engineering*. 99, pp. 1–1.

Torgo, L. 2010. Data Mining with R, learning with case studies Chapman and Hall/CRC. URL: http://www.dcc.fc.up.pt/~ltorgo/DataMiningWithR.

Wang, Z., Hope, R.M., Wang, Z., Ji, Q., Gray, W.D., 2012. Cross-subject workload classification with a hierarchical Bayes model. *NeuroImage, Neuroergonomics: The human brain in action and at work 59*, 64–69. doi:10.1016/j.neuroimage.2011.07.094.

Wijsman, J., Grundlehner, B., Liu, H., Penders, J., Hermens, H., 2013. Wearable Physiological Sensors Reflect Mental Stress State in Office-Like Situations, in: *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*. pp. 600–605. doi:10.1109/ACII.2013.105.

Williamon, A., Aufegger, L., Wasley, D., Looney, D., Mandic, D.P., 2013. Complexity of physiological responses decreases in high-stress musical performance. *Journal of The Royal Society Interface 10*. doi:10.1098/rsif.2013.0719.

Wilson, G. F.and Russell, C. A., 2003. Real-Time Assessment of Mental Workload Using Psychophysiological Measures and Artificial Neural Networks. *Human Factors, 45*, 635-643.

Yin, Z., Zhang, J., 2014. Operator functional state classification using least-square support vector machine based recursive feature elimination technique. *Computer Methods and Programs in Biomedicine 113*, 101–115. doi:10.1016/j.cmpb.2013.09.007.