A Composite Cognitive Workload Assessment System in Pilots Under Various Task Demands Using Ensemble Learning

Hyuk Oh[1(✉)], Bradley D. Hatfield[1,2], Kyle J. Jaquess[2], Li-Chuan Lo[2], Ying Ying Tan[1], Michael C. Prevost[5], Jessica M. Mohler[6], Hartley Postlethwaite[7], Jeremy C. Rietschel[8], Matthew W. Miller[9], Justin A. Blanco[7], Shuo Chen[4], and Rodolphe J. Gentili[1,2,3]

[1] Neuroscience and Cognitive Science Program, University of Maryland, College Park, MD 20742, USA
hyukoh@umd.edu
[2] Department of Kinesiology, University of Maryland, College Park, MD 20742, USA
[3] Maryland Robotics Center, University of Maryland, College Park, MD 20742, USA
[4] Department of Epidemiology and Biostatistics, University of Maryland, College Park, MD 20742, USA
[5] Naval Survival Training Institute, Pensacola, FL 32508, USA
[6] Midshipmen Development Center, United States Naval Academy, Annapolis, MD 21402, USA
[7] Electrical and Computer Engineering Department, United States Naval Academy, Annapolis, MD 21402, USA
[8] Maryland Exercise and Robotics Center of Excellence, VA Maryland Health Care System, Baltimore, MD 21201, USA
[9] School of Kinesiology, Auburn University, Auburn, AL 36849, USA

Abstract. The preservation of attentional resources under mental stress holds particular importance for the execution of effective performance. Specifically, the failure to conserve attentional resources could result in an overload of attentional capacity, the failure to execute critical brain processes, and suboptimal decision-making for effective motor performance. Therefore, assessment of attentional resources is particularly important for individuals such as pilots who must retain adequate attentional reserve to respond to unexpected events when executing their primary task. This study aims to devise an expert model to assess an operator's dynamic cognitive workload in a flight simulator under various levels of challenge. The results indicate that the operator's cognitive workload can be effectively predicted with combined classifiers of neurophysiological biomarkers, subjective assessments of perceived cognitive workload, and task performance. This work provides conceptual feasibility to develop a real-time cognitive state monitoring tool that facilitates adaptive human-computer interaction in operational environments.

Keywords: Attentional reserve · Mental workload · Simulated visuomotor task · Ensemble of classifiers

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC 2015, LNAI 9183, pp. 91–100, 2015.
DOI: 10.1007/978-3-319-20816-9_10

1 Introduction

In a variety of everyday situations, we observe that different people process identical stimuli in different ways. Specifically, we can focus our attentional resources on particular parts of a stimulus (e.g., spatial locations or shapes of a target object in visual scene) for detailed analysis and optimal decision-making, and overlook irrelevant aspects. Particularly in cognitive-motor performance studies, attention refers to the allocation of limited cognitive resources to execute a task [1], and our previous works have confirmed that it is consumed in proportion to task demand [2–4]. From these findings, three critical facets of attention and task demand can be inferred; (a) attention is a limited resource that can be focused on a single task or divided among several tasks, (b) individuals can shift their attentional focus between tasks having different requirements in terms of the size of the focus, and (c) cognitive-motor performance is sensitive to these differences including task demands.

It was suggested that attentional reserve (i.e., the unused portion of attentional capacity) is a main factor to measure mental workload (i.e., the used portion of attentional capacity) along with task demand [5]. Moreover, many studies have assessed mental workload by means of various metrics such as subjective ratings, the secondary-task paradigm, and psychophysiological measures (e.g., heart rate, galvanic skin response, evoked potentials, and pupil diameter) [2–4, 6, 7]. Overall, these results imply that the excessive mental workload from a failure to conserve adequate attentional reserve would result in an overload of attentional capacity, the failure to execute critical brain processes, and suboptimal decision-making for effective motor performance. As such, assessment of mental workload and attentional reserve is particularly important for individuals such as pilots who must retain adequate reserve to respond to unexpected events when executing their primary task. In this context, our previous work examined subjective report (Visual Analog Scale (VAS) and NASA Task Load Index (NASA-TLX)) as well as electroencephalographic (EEG) and electrocardiographic (ECG) biomarkers in conjunction with detailed monitoring of task-specific behavioral performance during a flight simulation task characterized by various task demands [4]. This result revealed that multiple metrics could index mental workload in ecologically valid situations.

Although this finding identified selective biomarkers to construct a composite metric sensitive to mental workload, there is still a need to further investigate a set of quantifiable features to assess the operator's cognitive workload in a wide variety of operational situations. However, this is very challenging as the data originates from various sources and possess different characteristics. Surprisingly, only a handful of studies have classified multimodal metrics derived from EEG, ECG, or pupillometry, according to the task demands in a pilot task (e.g., [8, 9]). Although interesting, their approaches ended up in studying individual classifiers to achieve higher accuracy rather than examining the relationship between features as well as classifiers.

Therefore, we propose to employ an advanced machine learning algorithm called *ensemble learning* to better understand how multiple metrics from various modalities correlate in feature space, and how the selected heterogeneous features relate to the operator's cognitive state. Ensemble learning is a process that creates and combines a set of independent expert classifiers to improve the prediction performance of a model

[10]. In other words, if data obtained from multiple sources contain complementary information (e.g., amplitudes and latencies of event-related potentials (ERP), standard deviation of the N-N interval (SDNN), the root mean square of the successive differences (RMSSD) of N-N intervals for heart rate variability (HRV), etc.), a proper fusion of such information can lead to improved accuracy of the prediction, even if the predictions from each individual data is less accurate. Thus, this study aims to devise an ensemble model that is able to select optimal features from multiple metrics for a more accurate classification according to three levels of task demand to determine an operator's current cognitive workload in ecologically valid tasks.

2 Methods

2.1 Data Acquisition

Subjects. Thirty-nine healthy volunteers (35 men and 4 women) between the ages of 19 and 24 years (mean and standard deviation age of 20.79 and 1.18), who were midshipmen in the United States Naval Academy (USNA), participated in this study. All subjects had received basic flight training covered in Aviation Preflight Indoctrination before participating in the study.

Apparatuses. Three systems were employed: (1) electrocortical and physiological data acquisition system, (2) a flight simulator cockpit and the Flight Data Recorder (FDR) known as one part of the black box, and (3) auditory stimuli with synchronous trigger delivery system. First, EEG and ECG recordings were accomplished using a single amplifier (g.USBamp®, g.tec medical engineering). Specifically, four active gel-free EEG electrodes (g.SAHARA electrodes®) were placed on the scalp along the frontal, frontocentral, central, and parietal midline sites (Fz, FCz, Cz, and Pz respectively) according to the 10–20 System. In addition, one-lead ECG electrode was placed below the 8th rib for basic heart monitoring. The system was grounded to the right mastoid and referenced to the left ear (A1); recording from the right ear (A2) was used for later EEG re-referencing purpose. Second, the simulator cockpit was equipped with a 22-inch widescreen LCD monitor, a computer with external stereo speakers, a Hands On Throttle-And-Stick set, and rudder pedals. Prepar3D® (v1.4, Lockheed Martin) was installed in the simulator computer, and a custom FDR was used to log operating conditions of the flight such as time, airspeed, and heading that will reflect the pilot's motor response and quality of flight operational performance. Finally, Presentation® (v18.1, Neurobehavioral Systems) delivered auditory probes to the subjects through earphones [2], and sent digital TTL pulses to the g.USBamp through a parallel port to mark stimulus-dependent synchronous triggers on the EEG and ECG data. The volume on the earphones was adjusted to a comfortable and yet audible level for each subject to ensure that the subjects could hear the engine and other mechanical noises played through the external speakers.

Scenarios and Task. Three scenarios were selected from the flight training program and adjusted with advice from naval aviators. Specifically, S1 was to keep straight and level flight, S2 was to repeat straight descending and climbing flight, and S3 was to

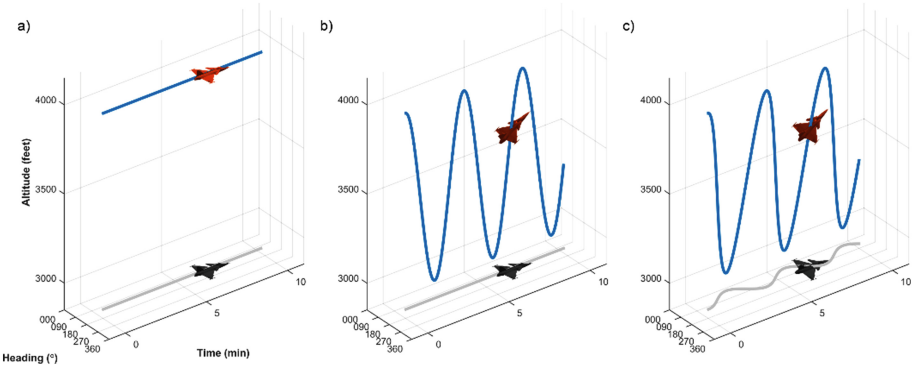

Fig. 1. A brief outline of three challenging scenarios; the red aircraft and blue line respectively depict the manned aircraft and its trajectory for 10 min, and they are represented by gray shadows. (a) S1 is to maintain straight and level flight. (b) S2 is to repeat straight descending and climbing flight. (c) S3 is to repeat diving in left turns and climbing in right turns.

Fig. 2. A flight simulation task; (a) EEG, ECG, and flight data were monitored while the subjects were engaged in the task. (b) the main instrument console in virtual cockpit contains all instrumentation and systems displays for the aircraft. The cockpit layout is based on the US air force and navy joint primary aircraft training system training documentation.

repeat diving in left turns and climbing in right turns (Fig. 1; [4, 11]). Thus, they represented relatively low, moderate, and high demand scenarios, respectively. Scenario sequence was counter-balanced. Participants controlled a single-engine turboprop aircraft (Beechcraft T-6A Texan II). They were instructed to use the primary flight controls (i.e., the control stick, the rudder pedals, and the throttle), without the use of secondary flight controls and other controls (e.g., rudder/elevator/aileron trim systems, etc.; Fig. 2). The flight was programmed to begin at 0900 virtual time, and at an initial altitude of 4000 feet.

Data Acquisition. Each participant sat in the simulator cockpit and was allowed 5 min of free practice along with exposure to the novel sounds. After the practice session, they were prepared for placement of the EEG and ECG sensors. The amplifier was calibrated, and all electrode impedances were maintained below 5 kΩ during data acquisition. The

signals were sampled at a rate of 512 Hz with an online Butterworth filter from 0.01 to 60 Hz. Participants were then assigned an initial scenario with relevant instructions. Each scenario was composed of a 1 min preparation period followed by a 10 min scenario. During each scenario, up to 30 stimuli were randomly presented with an inter-stimulus interval randomly ranging from 6 to 30 s, and all relevant flight data were sampled at a rate of 2 Hz. After the completion of each scenario, participants were provided the five VAS questions and the NASA-TLX survey to report their subjective experience. The same order of procedures was followed until all scenarios were completed.

2.2 Preprocessing

EEG. The re-referenced EEG were processed using an IIR filter with a 20-Hz low-pass cut-off frequency and 48-dB/octave roll-off. Next, each baseline of 1-s epochs that were time-locked to the auditory stimuli was corrected using the pre-stimulus interval (−100 to 0 ms). Those epochs retaining significant artifacts (e.g., eye-blink, etc.) were excluded. The remaining epochs were averaged per sensor for each scenario. The average ERP amplitudes were derived for the novelty P3 (P3a; 270 to 370 ms).

ECG. HRV was measured through the following methods along with average heart rate, which are appropriate for short 5 min samples from the middle section of the whole 10 min experiment: (1) SDNN, (2) RMSSD, (3) the low (LF; 0.04 to 0.15 Hz) and high frequency (HF; 0.15 to 0.4 Hz) ratio (LF/HF) of R-R intervals using a Welch's method.

FDR. Acceptable operational performance was defined as deviations from the tolerance limits of the target goals: at most ±200 feet of altitude, ±10 knots of airspeed, ±5° of heading, ±5° of bank angle, and ±500 fpm of ascent/descent rate. Each deviation was bounded above and below the acceptable decision boundary. The performance was calculated once a minute using the area under bounded curves, and normalized in the early (0 to 2 min), middle (4 to 6 min), and late (8 to 10 min) segments of the tests.

Subjective Response. All VAS and NASA-TLX variables were measured to range from 0 to 100, where greater value indicates the respondent felt relatively more efforts in the corresponding scenario.

2.3 Classification

First, attribute sets were constructed by joining preprocessed metrics. This allows each data to be expressed uniquely as a combination of basis vectors (i.e., selective attributes), which means that the attributes for the optimal classifier can better represent the relationship between the data and the task demands. Five individual classifiers were examined: classification trees (CTREE), k-nearest neighbors (kNN), quadratic discriminant analysis (QDA), naïve Bayes (NB), and error-correcting output codes using support vector machine (ECOC-SVM). For an ensemble of classifiers, bagging, boosting, stacking, and voting algorithms were scrutinized. Specifically, bagging derived the final prediction through a simple majority rule from multiple CTREEs.

Boosting combined weak CTREEs using a weighted majority rule, where each classifier was sequentially built as a better model than previous classifiers by considering misclassified observations. Unlike bagging and boosting, stacking employed a higher level model (CTREE) to combine five base learners (i.e., individual classifiers) rather than using one algebraic rule, and voting averaged the predictions of the five base learners. For each classifier, various tests were simulated to find the optimal parameters, and to determine the optimal classifier. Specifically, the optimal model was selected by taking the minimum balance (BAL) error (ε_{BAL}) to balance large biases due to a small sample size [12], which is a convex combination of resubstitution (RESB) and cross-validation (CV) errors. Lastly, the classifiers were assessed using the confusion matrix and the Receiver-Operating Characteristic (ROC) curve, and the rank of attributes was assessed through the ReliefF algorithm.

3 Results

Up to 117 samples (39 subjects in 3 scenarios) with 35 attributes (EEG, SDNN, Airspeed, etc.) from 4 metrics (P3a Amplitude, HRV, FDR, and Subjective Response) including missing data were used to train and test the classifiers.

3.1 Individual Classification

Each optimal classifier was constructed by exhaustively searching various possible parameters and CV settings (Table 1). According to our empirical tests, ECOC-SVM outperformed other classifiers in terms of RESB error (ε_{RESB}), but leave-one-out

Table 1. Optimal individual classifiers for each metric

Metric	Assessment	CTREE	kNN	QDA	NB	ECOC-SVM
P3a Amplitude	ε_{RESB}	0.2105	0.2807	0.5088	0.4912	0.0351
	ε_{CV}	0.6316	0.4211	0.7018	0.5789	0.5439
	ε_{BAL}	0.4211	0.3509	0.6053	0.5351	0.2895
	Accuracy	0.5000	0.7143	0.3750	0.5714	**0.8571**
HRV	ε_{RESB}	0.2464	0.4348	0.6667	0.4928	0.0290
	ε_{CV}	0.7826	0.5797	0.6812	0.8986	0.7681
	ε_{BAL}	0.5145	0.5072	0.6740	0.6957	0.3986
	Accuracy	0.6667	0.5000	0.3333	0.3333	**0.7500**
FDR	ε_{RESB}	0.0601	0.1021	0.2583	0.2763	0.0601
	ε_{CV}	0.1862	0.1411	0.2583	0.2973	0.2823
	ε_{BAL}	0.1232	0.1216	0.2583	0.2868	0.1712
	Accuracy	0.8000	**0.9333**	0.8333	0.8333	0.8667
Subjective Response	ε_{RESB}	0.1282	0.2478	0.2112	0.2821	0.0769
	ε_{CV}	0.3590	0.2743	0.4431	0.3419	0.3846
	ε_{BAL}	0.2436	0.2611	0.3272	0.3120	0.2308
	Accuracy	0.7333	0.7333	0.6000	0.8000	**0.8667**

cross-validated kNN surpassed all the other classifiers regarding CV error (ε_{CV}). Thus, no single classifier can unvaryingly outperform the others over all datasets. Each algorithm has its unique strengths, so it is difficult to decide which method is most appropriate on each dataset. However, the results revealed that εBAL could be a reliable indicator to select the optimal model, although the predictive accuracy is not proportional.

The predictive accuracy rates were 85.71 % for P3a Amplitude; 75.00 % for HRV; 93.33 % for FDR; 86.67 % for Subjective Response. However, to exaggerate, the three weaker classifiers have no contribution to assessing overall mental workload, because the assessment will be still robust with accuracy of 93.33 % even if three less accurate classifiers and corresponding data are excluded from the system. On the other hand, it will be impractical to mix all mutually unrelated metrics into one container to construct a single individual classifier that generates precise and accurate results for all datasets, because it will increase the overall complexity concerning system architecture as well as data representation, classification, and interpretation.

3.2 Ensemble Classification

There is no clear and comprehensive picture of which ensemble methods are optimal. In Table 2, bagging, boosting, and stacking respectively reached an accuracy of 88.0, 95.7, and 94.0 %, while voting had the highest 97.44 %.

The bagging showed that the ensemble learning did not guarantee good performance all the time, although it was still reasonable. Possible reasons may include several missing values on some variables (e.g., certain subjects skipped some VASs), small data sizes, unexpectedly correlated classifiers, noise in samples, and suboptimal parameters. Another critical reason may be due to ill-conditioned data, because some subjects were overwhelmed during high demanding scenarios and had given up halfway through the task. However, the boosting, stacking, and voting showed very reliable performance as expected. In particular, stacking and voting could be easily extensible through a hierarchical structure particularly when any new feature set is included in the future to produce more reliable results.

Table 2. Optimal ensemble classifiers and assessment measurements

a) Bagging

Accuracy	0.880	True Class			Precision	ROC
Kappa	0.821	S1	S2	S3		
Predictive Class	S1	33	4	2	0.846	0.927
	S2	6	31	2	0.886	0.897
	S3	0	0	39	0.907	0.958
Recall		0.846	0.795	1.000		

b) Boosting (AdaBoost)

Accuracy	0.957	True Class			Precision	ROC
Kappa	0.936	S1	S2	S3		
Predictive Class	S1	37	0	2	0.974	0.992
	S2	1	36	2	1.000	0.984
	S3	0	0	39	0.907	0.989
Recall		0.949	0.923	1.000		

c) Stacking

Accuracy	0.940	True Class			Precision	ROC
Kappa	0.910	S1	S2	S3		
Predictive Class	S1	37	2	0	0.974	0.996
	S2	1	36	2	0.900	0.984
	S3	0	2	37	0.949	0.974
Recall		0.949	0.923	0.949		

d) Voting

Accuracy	0.974	True Class			Precision	ROC
Kappa	0.962	S1	S2	S3		
Predictive Class	S1	38	0	1	0.974	0.995
	S2	1	37	1	1.000	0.999
	S3	0	0	39	0.951	1.000
Recall		0.974	0.949	1.000		

Table 3. Importance of selective attributes by means of ReliefF algorithm using 8-fold CV

Average Merit	Average Rank	Attribute	Average Merit	Average Rank	Attribute
0.1813	2.37	Heading	0.0190	23.40	P3a Amplitude on Pz
0.1073	7.67	Altitude	0.0140	27.10	P3 Amplitude on Cz
0.0430	17.66	VAS	0.0110	27.40	RMSSD
0.0290	18.90	P3a Amplitude on Fz	0.0130	27.40	P3a Amplitude on FCz
0.0303	20.23	Vertical Speed	0.0100	28.10	SDNN
0.0247	21.23	Airspeed	0.0080	28.30	LF/HF
0.0248	22.87	NASA-TLX	0.0070	29.40	HR

ReliefF evaluator showed that most of the attributes related to FDR and self-reports better represent task demands in the optimal ensemble classifier (Table 3). It is no wonder that both features are strongly related with task demands, because FDR directly reflects how the operators are performing to achieve a given task, and subjective responses quantify their internal status about the task demands. Interestingly, this result revealed that the P3a Amplitudes (7.5 %) and RMSSD (1.1 %) had a fair amount of contribution to the assessment of the operator's mental workload even considering strong attributes such as FDR (34.36 %) and self-reports (6.78 %). In addition, this revealed that there is a proportional relationship between the importance of attributes and their statistical significance according to our previous work [4].

4 Discussion

This study proposed a novel approach to assess the operator's mental workload under various task demands in ecologically valid situations. Although other research groups have reported the feasibility of mental workload assessment with specific classifiers for each metric, they have mainly focused on individual-level analysis instead of providing results for combined classification [9, 13]. Moreover, only a few studies have examined even individual-level classifiers using multimodal data from operational environments (e.g., [9, 14]). Thus, this study complements previous efforts in that both individual and ensemble learning of multiple classifiers were employed to examine the relationship between mental workload and various feature vectors in an ecologically valid task. Particularly, compared to other methods, while providing accurate and reliable classification accuracy, our approach is able to handle diverse new datasets (e.g., functional near-infrared spectroscopy) and other task specific performance measurements by combining them with existing classification models considering the characteristics of ensemble learning.

We showed that mental workload is a predictable variable with high accuracy if ensemble classifiers are optimally configured. Specifically, high classification accuracy can be achieved when each expert detects distinct but common directional patterns from each feature set, and the final arbiter of classification makes a decision considering a majority vote from each expert. Moreover, the ReliefF indicates that even

statistically non-significant metrics (e.g., HRV in our previous study [4]) could contribute to constructing the correct classification if they are discriminable.

Our current research efforts were limited to classification, where the model could predict categorical dependent variables. This method could not predict graded cognitive states, because the task demand and mental workload are intrinsically continuous. Thus, ensemble regression is worth considering in the future. However, constructing such approximation models require highly complex computation, but such a burden can be alleviated by constructing surrogate models. Other possible future work is to pursue more practical approaches. For example, current ERP analysis was highly dependent on the novel sounds, which will be inappropriate when cognitive tasks include auditory stimulus or a silent environment such that the introduction of extraneous sounds could be destructive. One solution is to employ the eye-tracking technology to extract the fixation-related potential (FRP) on the areas of interest, while another solution is to use other signal processing methods that do not rely on ERPs as suggested by our colleagues at the USNA [11].

In summary, the results revealed that both individual and combined classifiers could effectively assess properly constructed feature sets that were extracted from multimodal data. Particularly, ensemble classifiers are expected to outperform individual classifiers, because a single strong referee of classification could merge information collected from multiple weaker experts. As a long-term goal, this work provides conceptual feasibility to develop a real-time cognitive state monitoring tool that facilitates adaptive human-computer interaction in operational environments.

Acknowledgement. The authors would like to express their appreciation for the guidance and supports provided by Kenneth T. Ham (Captain, USN) who was a naval astronaut and the chair of the Aerospace Engineering Department at the USNA, and all student aviators who were midshipmen at the USNA and volunteered to participate in the study. In addition, this research was supported by the Lockheed Martin Corporation (Bethesda, MD, USA) under grant 4-321830.

References

1. Kahneman, D.: Attention and Effort. Prentice-Hall, Englewood Cliffs (1973)
2. Miller, M.W., Rietschel, J.C., McDonald, C.G., Hatfield, B.D.: A novel approach to the physiological measurement of mental workload. Int. J. Psychophysiol. **80**, 75–78 (2011)
3. Rietschel, J.C., Miller, M.W., Gentili, R.J., Goodman, R.N., McDonald, C.G., Hatfield, B. D.: Cerebral-cortical networking and activation increase as a function of cognitive-motor task difficulty. Biol. Psychol. **90**, 127–133 (2012)
4. Gentili, R.J., Rietschel, J.C., Jaquess, K.J., Lo, L.-C., Prevost, M.C., Miller, M.W., Mohler, J. M., Oh, H., Tan, Y.Y., Hatfield, B.D.: Brain biomarkers based assessment of cognitive workload in pilots under various task demands. In: 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 5860–5863. IEEE, Chicago (2014)
5. Kantowitz, B.H.: Mental workload. In: Hancock, P.A. (ed.) Human Factors Psychology, pp. 81–121. Elsevier, Amsterdam (1987)
6. Hankins, T.C., Wilson, G.F.: A comparison of heart rate, eye activity, EEG and subjective measures of pilot mental workload during flight. Aviat. Space Env. Med. **69**, 360–367 (1998)

7. Kok, A.: Event-related-potential (ERP) reflections of mental resources: a review and synthesis. Biol. Psychol. **45**, 19–56 (1997)
8. Wilson, G.F., Russell, C.A.: Real-time assessment of mental workload using psychophysiological measures and artificial neural networks. Hum. Factors **45**, 635–643 (2003)
9. Noel, J.B., Bauer, K.W.J., Lanning, J.W.: Improving pilot mental workload classification through feature exploitation and combination: a feasibility study. Comput. Oper. Res. **32**, 2713–2730 (2005)
10. Dietterich, T.G.: Ensemble methods in machine learning. In: Kittler, J., Roli, F. (eds.) MCS 2000. LNCS, vol. 1857, pp. 1–15. Springer, Heidelberg (2000)
11. Johnson, M.K., Blanco, J.A., Gentili, R.J., Jaquess, K.J., Oh, H., Hatfield, B.D.: Probe-independent EEG assessment of mental workload in pilots. In: 7th International IEEE EMBS Conference on Neural Engineering. IEEE, Montpellier (2015)
12. Raudys, S.J., Jain, A.K.: Small sample size effects in statistical pattern recognition: recommendations for practitioners. IEEE Trans. Pattern Anal. Mach. Intell. **13**, 252–264 (1991)
13. Henelius, A., Hirvonen, K., Holm, A., Korpela, J., Müller, K.: Mental workload classification using heart rate metrics. In: Proceedings of 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2009, pp. 1836–1839 (2009)
14. Kohlmorgen, J., Dornhege, G., Braun, M.L., Blankertz, B., Müller, K.-R., Curio, G., Hagemann, K., Bruns, A., Schrauf, M., Kincses, W.E.: Improving human performance in a real operating environment through real-time mental workload detection. In: Dornhege, G., del Millán, J.R., Hinterberger, T., McFarland, D., Müller, K.-R. (eds.) Toward Brain-Computer Interfacing, pp. 409–422. MIT Press, Cambridge (2007)