This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2016.2628702, IEEE Transactions on Cognitive and Developmental Systems

Quantifying Cognitive Workload in Simulated Flight Using Passive, Dry EEG Measurements

Justin A. Blanco*, *Senior Member, IEEE*, Michael K. Johnson*, Kyle J. Jaquess, Hyuk Oh, Li-Chuan Lo, Rodolphe J. Gentili, and Bradley D. Hatfield

***Abstract*— A reliable method for quantifying cognitive workload in pilots could find uses in flight training and scheduling, cockpit design, and improving flight safety. Many proposed methods for monitoring cognitive workload in this population rely on measuring physiological responses to externally delivered probe stimuli and/or use traditional gel-based electroencephalography (EEG) sensors. Here we develop passive, probe-independent algorithms for classifying three levels of flight task complexity based on 4-channel, gel-free EEG during simulated flight. Using a library of 168 input features drawn from different data science application domains, we evaluated 13 different classifiers, using a nested 10-fold cross-validation procedure to estimate generalization performance. The best subsets of features yielded a median classification accuracy of 90.17% across subjects, with perfect accuracy in 1 subject and greater than 75% in 16 of 21 subjects. Though EEG line length and linear discriminant analysis were generally among the most effective features and classifiers, respectively, we find that to maximize prediction accuracy, feature set-classifier combinations should be individualized. No single channel proved more valuable than another in predicting flight task complexity, but combining EEG features across channels maintained or improved performance in 81% of subjects.**

***Index Terms*—Cognitive Workload, electroencephalography, pilots, machine learning**

I. Introduction

COGNITIVE workload can be characterized as a ratio between task complexity and a person's cognitive capacity to meet task demands [1]. This description captures the intuitive idea that cognitive workload depends both on external factors such as the objective difficulty of required tasks and individual factors such as a person's skills and past experiences. There is a vast and growing body of research focused on developing quantitative methods to assess cognitive workload for a variety of purposes, for example to study the cognitive resiliency of people in high stress environments. Metrics derived from physiological signals such as heart and respiration rate; brain activity; blood pressure; galvanic skin response; pupil diameter; and eye-blink and gaze have been investigated as biomarkers of cognitive workload (e.g., [2]–[5] and many others). These signals have demonstrated associations with cognitive workload and have in some cases been used, either alone or in combination, to distinguish cognitive workload levels with accuracies significantly better than chance. Still, there remains a need to propose systems that are sensitive enough for cognitive workload monitoring outside of the laboratory while simultaneously being easy and comfortable to use, portable, and durable. In that regard, with recent improvements in the ease-of-use, reliability, and costs of portable electroencephalography (EEG) systems, there is increasing interest in using and optimizing brain signal-related measurements of cognitive workload, potentially even with consumer grade equipment [6]. It is reasonable to expect that EEG offers a more direct assay of cognitive workload than other physiological biomarkers because of the proximity of EEG sensors to the neural substrates of cognitive stress [7]. Furthermore, previous work has shown better performance for EEG than eye and peripheral measurements in a binary mental workload classification task [8]. Particularly in space, cost, or other resource-constrained scenarios where it may be impractical to instrument a subject with multiple-modality physiological sensors, EEG may be the preferred choice for cognitive workload monitoring – especially as gel-free and non-contact electrode systems capable of being embedded in standard helmets or caps [9] become more widely available.

A recent study employed EEG to assess cognitive workload in pilots while performing an aircraft flying task under three different levels of challenge [10]. In contrast with previous similar investigations, this study employed only few dry (i.e., no conductive gel) EEG electrodes to derive EEG power combined with an event-related potential (ERP) component, the P300 [11]. The P300 is a voltage deflection that is one component of a well-studied brain response often referred to as the oddball potential. It occurs approximately 300ms following the introduction of a rare stimulus amidst a series of higher-probability stimuli. In this particular experiment, the P300 was elicited with a relatively new auditory probe technique that confirmed and extended previous laboratory work by revealing that P300 amplitude varies inversely with task complexity [12].

Although this and similar techniques that measure ERPs using auditory stimuli have been successful in evaluating cognitive workload in controlled settings, this approach may be

Manuscript received February 16, 2016. This work was supported in part by Office of Naval Research awards N001614WX30023 and N001613WX20992 and by Lockheed Martin Corporation grant 4-321830.

J.A. Blanco is with the Department of Electrical and Computer Engineering, United States Naval Academy, Annapolis, MD 21402 USA (email: blanco@usna.edu).

M.K. Johnson is with the Department of Computer Science, Stanford University, Stanford, CA 94305 USA (email: michael.johnson@stanford.edu).

K.J. Jaquess, H. Oh, L. Lo, R.J. Gentili, and B.D. Hatfield are with the Department of Kinesiology, University of Maryland, College Park, MD 20742 USA (emails: rodolphe@umd.edu; bhatfiel@umd.edu).

*J.A. Blanco and M.K. Johnson contributed equally to this work.

less practical when required cognitive tasks themselves involve a strong auditory component and hence the introduction of extraneous sounds could be disruptive. Therefore, an EEG-based metric of cognitive workload that exploits information from non-evoked "background" neural activity is desired [13], [14].

The goal of this research was to develop a background EEG-based algorithm to classify multiple levels of aircraft pilot-related task complexity using a small number of dry electrodes without relying on ERPs. We used flight simulator tasks of varying complexity as our experimental paradigm because we are predominantly concerned with understanding the response of aircraft pilots to the cognitive demands imposed by their missions. Furthermore, pilots are usually in constant radio or intercom communications via headset during flight, a scenario that is be particularly well suited to an index of cognitive workload not based on ERPs.

Several studies have examined the relationship between background EEG and cognitive workload in aviation environments with a range of results [15]. For example, Hankins and Wilson [2] found an increase in theta band EEG power during segments of a flight-based task that required mental calculations. Dussault and colleagues found that delta band activity also increased with cognitive workload, while beta and gamma band activity decreased [16], [17]. However, beta and gamma activity have also been shown to have positive relationships with cognitive workload [18], [19]. Research by Berka and colleagues [20] utilized a cognitive workload metric based upon a weighted linear combination of various spectral bands gathered from a collection of bipolar EEG sites and showed that the EEG-based metric had a positive relationship with cognitive workload.

Beyond showing an associative relationship between cognitive workload and EEG or other physiological metrics, several researchers have used machine-learning approaches to predict or classify cognitive workload. For example, a k-NN algorithm was employed to estimate cognitive workload on the basis of the prefrontal blood oxygenation changes [21]. Other studies have shown that either Artificial Neural Networks (ANNs) [22] or Linear Discriminant Analysis (LDA) [23] based classifiers can monitor workload from EEG power spectra in real-time. In addition, some classification algorithms such as ANNs [24] and ensemble methods [14] have been applied to data sets combining psychophysiological measures such as EEG, cardiovascular, and ocular signals to predict pilots' cognitive workload.

In this paper, we use machine-learning algorithms to classify cognitive workload in a simulated flight environment. The work builds on and is distinguished from other groups' work on cognitive workload in aviation environments in that it simultaneously: 1) uses data from only a small number (four) of EEG sensors; 2) uses dry (gel-free) electrodes; 3) requires only non-evoked background EEG data; 4) analyses EEG segments of durations amenable to real-time processing; and 5) successfully classifies more than two levels of cognitive workload. We believe that all five of these aspects are critical steps toward realizing a practical cognitive workload monitoring device for pilots in the cockpit. In a preliminary study [13] using a subset of the subject pool examined here, we reported the performance of a classifier that used only EEG spectral band power measurements as input features. Estimates of the EEG power in discrete frequency ranges with known clinical significance commonly appear as input features in the brain-machine interface literature [15], [25]–[27], including other work on cognitive workload measurement [8], [28]–[33], as previously discussed. Our results using only alpha (8-13 Hz), beta (13-30 Hz), and gamma (30-40 Hz) power as classifier inputs were modest: sixty-eight percent (13 of 19) of subjects had at least one EEG channel yielding classification accuracy above 50%, with chance performance at 33%.

The present study extends the latter work in several important ways. We demonstrate that cognitive workload classification using background EEG collected from dry sensors is substantially improved by expanding the set of input features and supervised learning algorithms. We then demonstrate further improvements by employing state-of-the art feature subset selection and dimensionality reduction techniques. Finally, we use subjects' behavioral performance data (i.e., measures of how adeptly they perform the experimental flight tasks) to gain insight into the observed relationship between background EEG and cognitive workload. In particular, we provide evidence supporting the idea that while increasing task complexity is associated with decreasing behavioral performance, these two variables are not interchangeable. Cognitive workload increases that are detectable by EEG monitoring but do not have concomitant measureable behavioral performance decreases are particularly important in that they may reflect an individual's compromised ability to mount a successful response to new, emergent tasks which can arise while attending to a primary task.

II. Method

A. Data Acquisition

EEG data were collected from 21 United States Naval Academy (Annapolis, MD, USA) midshipmen between the ages of 19 and 23, all with basic flight training, while they performed visuo-motor tasks in a flight simulator (Prepar3D® v1.4, Lockheed Martin Corporation, Orlando, FL, USA) under three levels of task complexity. The fact that subjects had comparable previous flight experience served as a partial control for individual differences in the ability to perform the experimental tasks, and we therefore used task complexity as a proxy for cognitive workload. The three flight tasks were selected from predefined flight training exercises developed with advice from experienced United States Navy pilots and distinguished in challenge-level by differences in weather intensity and mission requirements. Specifically, the three tasks were: 1) *Easy*: maintain the aircraft's altitude (4000 ft), heading (180º), and airspeed (180 kn). The weather was defined by no clouds, precipitation, or wind, and unlimited visibility; 2) *Medium*: maintain the aircraft's heading (180º), airspeed (180 kn), and a "wings-level" attitude, while continuously making altitude changes between 4000 and 3000 ft, with ascent and descent rates of 1000 feet per minute (fpm). The sky was completely overcast (1/16 mi of visibility), but there was no precipitation and no wind; and 3) *Hard*: maintain the aircraft's airspeed (180 kn), while changing heading between 180º and 090º at a 15º angle of bank, ascending while turning right and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2016.2628702, IEEE Transactions on Cognitive and Developmental Systems

descending while turning left at 1000 fpm. The sky was completely overcast as in the *Medium* task, with no precipitation, but with the presence of a moderate (16 kn) easterly wind.

For each subject, one trial per task complexity was conducted in random order, consisting of a 1-minute setup period followed by a 10-minute flight segment. Additionally, for use in a separate analysis, audible stimuli were administered to participants via ear-bud speakers with random inter-stimulus intervals between 6 and 30 seconds to evoke the P300 response. Since the goal of this work was to analyze background EEG only, data surrounding the auditory stimuli were excluded as described below. Fig. 1 illustrates the experimental setup.

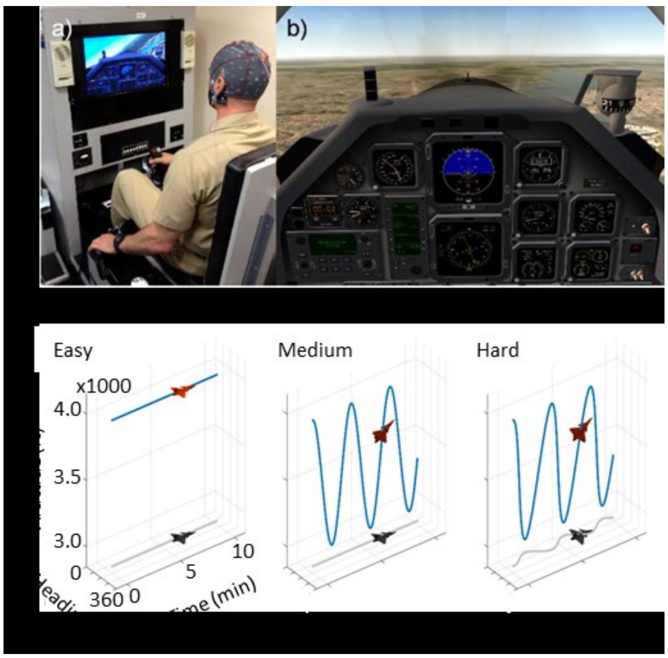

Fig. 1 a) Flight simulator setup: subjects operated a simulated T-6A Texan II SP2 United States Navy aircraft using a control stick, throttle, and rudder pedals; b) instrumentation and system status displays for the aircraft's virtual cockpit; c) visualization of the three experimental flight tasks: *Easy* - maintain straight and level flight; *Medium*- repeatedly ascend and descend; and *Hard* - ascend and descend while making left and right turns. All three plots have identical axes, including scaling. (Adapted from [14]).

Four active, gel-free electrodes were used to measure scalp EEG signals from sites along the frontal (Fz), fronto-central (FCz), central (Cz), and parietal (Pz) midline, based on the International 10-20 System [34]. The EEG cap was connected to an amplifier with an online bandpass filter from 0.01 to 60 Hz (g.USBamp®, g.tec medical engineering GmbH, Schiedlberg, Austria) through a driver-interface box (g.SAHARAsys®, g.tec medical engineering GmbH, Schiedlberg, Austria). Electrode impedances were maintained below 5 kOhm during data acquisition. The right mastoid was used as ground for the system and the left ear as the hardware reference. Data were also collected from the right ear for later re-referencing. EEG were sampled at a rate of 512 Hz, re-referenced in an EEG analysis software (BrainVision Analyzer 2, Brain Products GmbH, Munich, Germany) to an average-ear montage, and digitally lowpass filtered (in forward and reverse to give zero phase response) with a Butterworth filter with a cutoff frequency of 50 Hz and 48 dB/octave rolloff.

Prior to analysis, EEG data were visually inspected in the time-domain for the presence of eye-blink, eye-movement, muscle activity, cardiac, and respiratory artifacts, such as those described in [35]. Any segments containing the patterns associated with these artifacts or exhibiting a baseline drift of greater than ~50μV were excluded from analysis (31.18% of all data). In addition, data within a 600 ms window following the onset of each auditory stimulus were excluded prior to analysis in order to eliminate auditory-evoked responses.

B. Feature Extraction

The data set was segmented into 1-second (512-sample) epochs for analysis, based primarily on a desire to have no more than one second of lag in an envisioned online monitoring system. Table I shows the number of *Easy*, *Medium*, and *Hard* task EEG segments analyzed for each subject.

Table I. Raw data summary

1	(347,354,252)	**8**	(117, 27, 25)	**15**	(321,26,291)
2	(164,220,269)	**9**	(18, 52, 42)	**16**	(160,255,192)
3	(307,339,366)	**10**	(374,64,271)	**17**	(310,279,264)
4	(174,226,215)	**11**	(222,259,232)	**18**	(355,418,360)
5	(292,193,148)	**12**	(403,367,394)	**19**	(100,126,56)
6	(251,168,159)	**13**	(268,322,290)	**20**	(421,383,313)
7	(337,368,330)	**14**	(221,317,257)	**21**	(309,278,258)

Number of EEG segments analyzed per subject (1-21) for each level of task complexity (*Easy*, *Medium*, *Hard*).

Any linear trends were removed from each segment by subtracting the least squares line of best fit. Feature vectors were then formed for each 1-second epoch by concatenating various subsets (details in Section IIIB) of 672 EEG measurements – 168 different measurements on each of 4 channels. The features collectively represent information extracted from the time, frequency, and wavelet domains, as well as parameters derived from information theory, nonlinear dynamics, and fractal analysis. A detailed list of features and their associated references are contained in the Appendix.

C. Supervised Learning

Thirteen different classifiers were trained to predict task complexity based on EEG input features. Each individual observation classified was a feature-vector describing a 1-second EEG segment recorded from an individual subject in an *Easy*, *Medium*, or *Hard* simulated flight task. Broadly, the classification techniques used were: 1) k-Nearest Neighbors (kNN); 2) Discriminant Analysis; 3) Naïve Bayes; 4) Decision Trees (single, boosted, and Random Forests); and 5) Support Vector Machines (SVM). Support Vector machine classifiers, which are fundamentally binary, were made multi-way using the one-versus-all method [36]. Section A in the Appendix lists all classifiers and the values of their key properties and parameters.

Test set error was estimated using 10-fold cross-validation, with percent correct classification used as the performance metric. We compare classification performance on a

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2016.2628702, IEEE Transactions on Cognitive and Developmental Systems

per-channel basis using all 168 features (Section IIIA) to that obtained after performing feature subset selection and dimensionality reduction (Sections IIIB and IIIC), and by combining features across EEG channels before performing feature selection (Section IIID). Finally, in Section IV, we examine the relationship between subjects' behavioral task performance (i.e., their degree of success in executing the *Easy*, *Medium*, and *Hard* flight tasks) and our classification results.

III. Classification Results

A. All-Features Analysis

Because we were interested in exploring a minimal set of hardware requirements for reliable cognitive workload classification – and in particular whether good accuracy could be achieved using only a single electrode – our first analysis examined performance on an individual channel basis. Fig. 2 (top plot) displays, for each channel, the distribution of classification accuracies across all subjects for the best performing classifier using all 168 features as inputs. Accuracies above 50% (chance is 33%) were obtained for all subjects on at least one channel. The lowest classification accuracy was 45.91%, while the highest was 100%. Differences in best classifier performance among channels were not significant (p = 0.583, Friedman test). Of the 84 total (21 subjexts x 4 channels) best classifiers found, the most frquently occuring was LDA (33.3% of cases).

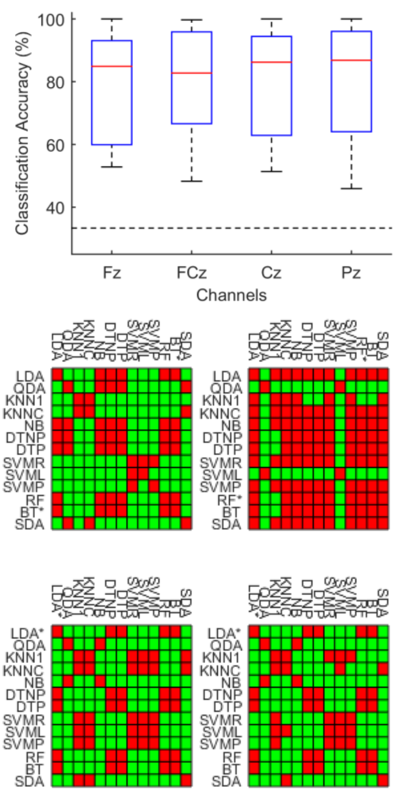

Fig. 2. Top: Each box-and-whisker plot represents the distribution of classification accuracies across all subjects for the best performing classifier using all 168 features. Red lines indicate the median of each distribution; the blue boxes span the first to third quartiles; and the whiskers extend to the extreme values. The single horizontal dotted line represents chance performance using a null model of equiprobable classes. Bottom: symmetric matrices summarizing the results of statistical significance testing for differences in classification accuracy between all 78 (13-choose-2) possible pairs of classifiers tested. Data shown are from four randomly selected subjects (clockwise from upper left: Subject 3, Subject 8, Subject 17, and Subject 20) and a single representative channel (FCz). Green cells represent a statistically significant difference in classification accuracy at the alpha level of 0.05 after Bonferroni correction for multiple comparisons. The best performing classifier for each of the four subject-channels shown is indicated with an asterisk. Classifier abbreviations can be found in the Appendix.

Within each channel, we used Friedman's test followed by post-hoc Bonferroni-corrected multiple pairwise comparisons to test whether observed performance differences among the classifiers were significant at the 0.05 alpha level. While the best classifier found was often statistically significantly superior to a majority of the others tested, it was frequently the case that a smaller subset of classifiers showed competitive performance with the best classifier. Most commonly this performance equivalence was among LDA and ensemble methods, particularly Boosted Trees (BT), and Random Forests (RF), as illustrated in Fig 2 (bottom plot).

Despite the improvement over the accuracies reported in [13] that was achieved by increasing the size of the input feature set, there remained considerable variability in classifier performance across subjects: the mean (over channels) interquartile range in fig. 2, for example, is 31.49.

B. Dimensionality Reduction

To mitigate overfitting, particularly in three subjects who had a relatively small number of observations (i.e., subjects 8, 9, and 19; see Table I), and to reduce the negative impacts on classification accuracy of any features that might be redundant or unrelated to cognitive workload, we separately performed both feature subset selection and Principal Components Analysis (PCA). For the former, since exhaustive search of the 2^{168} possible feature subsets to determine the optimal classification model for each channel is intractable, we used forward stepwise feature selection, which considers a much smaller search space of candidate models (at most $1+p(p+1)/2$, with p the number of features; 14197 models for p = 168 in our case).

Our forward stepwise feature selection procedure began with a model containing no features and, on each iteration, added to the model the single feature giving the greatest additional improvement in performance, measured as 10-fold cross-validated accuracy. The procedure halted when no addition of a single feature would yield an improvement in the performance criterion. Fig. 3 shows how the number of features included in the model affected classification performance in one particular subject (subject 2) who had relatively poor classification accuracy but a wide dynamic range of accuracies over the set of models considered.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2016.2628702, IEEE Transactions on Cognitive and Developmental Systems

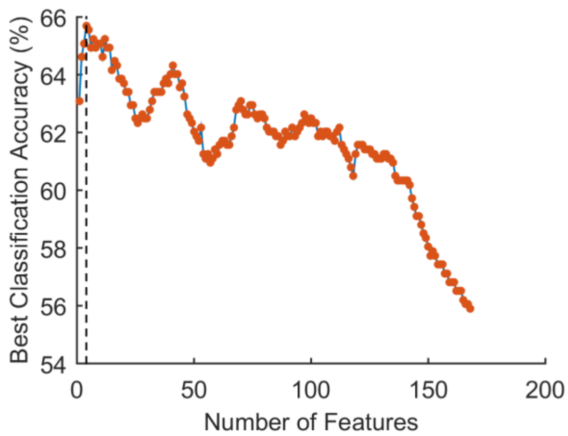

Fig. 3. Example illustrating forward feature selection. The vertical dotted line represents the number of features corresponding to the best performing subset as described in the text. The plot shows how classification performance evolves as the number of features increases to the maximum of 168 in this subject (Subject 2) and channel (FCz). In practice, to ease the computational burden we use the halting criterion described in the text, so the feature selection algorithm stops at the vertical line.

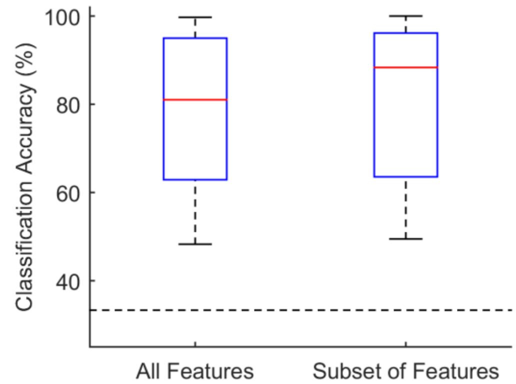

Fig. 4. Left: boxplot showing the distribution of best classification accuracies across subjects using all features as the input vector; Right: boxplot showing the distribution of best classification accuracies across subjects using the highest-performing subset of features as the input vector. Both distributions are for channel FCz. The dotted line represents chance performance using a null model of equiprobable classes. The Wilcoxon Signed Rank test rejects the null hypothesis that the difference between paired samples has zero median (Bonferroni corrected $p = 0.008$), indicating a statistically significant improvement in classifier performance by using a subset of features rather than all features.

In order to estimate and compare generalization performance across different classifiers – each trained using its own best feature subset determined by forward feature selection – we used a nested 10-fold cross-validation procedure [37] that ensured test samples used in the outer validation loop were not previously seen in model selection. Cross-validation was done on a per-subject basis using randomly generated folds comprised of a mixture of *Easy*, *Medium*, and *Hard* 1-second segments. To maintain computational tractability in the face of the nested cross-validation procedure, the five classifiers with the most computationally intensive implementations (SVML, SVMP, RF, BT, and SDA; see Appendix for all classifier abbreviations) were excluded from this feature selection analysis. All computations were carried out on a MATLAB (MathWorks, Natick, MA) compute cluster (96-worker job; compute nodes with four 12-core AMD Opteron 6174 processors at 2.2 GHz).

With feature selection, Linear Discriminant Analysis was again most frequently the best performing classifier, representing 42.86% (36 of 84) of all best classifiers found across subjects and channels. Performance among channels was not significantly different with feature selection ($p = 0.655$, Friedman test), similar to fig. 2. However, when comparing classifier performance with and without feature selection by channel, a significant performance improvement was seen with feature selection on all channels except Cz after multiple comparisons were accounted for (Wilcoxon Signed Rank test, Bonferroni corrected $p < 0.05$ for channels Fz, FCz, and Pz). Fig. 4 shows the distribution across subjects of best classification accuracies with and without feature selection for channel FCz.

Forward feature selection performs a greedy, guided search through the space of possible feature subsets and is inherently vulnerable to finding a local rather than global maximum of the classification accuracy objective. As a potential alternative to forward feature selection for dimensionality reduction we performed Principal Components Analysis [38] on the data, after normalizing each feature by subtracting its mean and dividing by its standard deviation. Principal Components Analysis finds the directions in input space along which the projected data are most variable, subject to the constraint that each successive direction found is orthogonal to the preceding one. The resultant projections onto these directions can be viewed as new input features that are linear combinations of the originals. To achieve dimensionality reduction, we retained the number of components required to preserve 99.9% of the total data variance, resulting in mean dimension reduction of 48.5%. Median classification performance using PCA was inferior to that of feature selection on all channels: the average median classification accuracy was 5.39% lower for PCA. However, this trend reached statistical significance on only one channel (FCz; Wilcoxon Signed Rank test, Bonferroni corrected $p = 0.039$).

C. Feature Importance

Although the nested cross-validation procedure we used for feature selection asymptotically yields an approximately unbiased estimate of true test error [37], because feature selection occurs in the inner loop, it does not in general identify a single unique best subset for each classifier. Therefore, in order to gain further insight into the relative importance of each feature for predicting task complexity, we conducted two separate, additional analyses.

In the first analysis, we used LDA as a prototype classifier to determine a unique best feature subset using forward selection. Linear Discriminant Analysis was used because it was the best performing classifier in 33.3% of cases during testing with all features and requires less computational time per iteration than

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2016.2628702, IEEE Transactions on Cognitive and Developmental Systems

other classifiers in our set. Since the final subset selected can depend on the random partitions of the data made during 10-fold cross validation throughout the progression of the feature selection algorithm, forward feature selection was performed 250 times for each combination of subject and channel, yielding 21×4×250 = 21,000 iterations. Table II lists the top five features, ordered by the percentage of iterations on which the feature was among the optimal subset selected. Time-series line length and number of peaks were the two most frequently selected features at 80.45% and 55.67%, of iterations respectively, considerably higher rates than any other features. This reflects both the relatively low effective dimensionality of the input space and the heterogeneity across subjects and channels with respect to which features have predictive value.

Table II. Forward Feature Selection Statistics

Feature Name	Selection Freq. (% of Iterates)	Avg. Rank in Subset
LTS	80.45	1.31
PTS	55.67	2.31
MTEF	15.02	2.38
HSD	14.00	3.70
MTEFZ	12.38	2.67

HSD: Hurst Exponent, Discrete Second-order Derivative; LTS: Line Length, Time Series; MTEF: Mean, Teager Energy Signal, Frequency Modulated Component; MTEFZ: Mean, Teager Energy Signal, Frequency Modulated Component, Z-Score; PTS: Number of Peaks, Time Series.

Fig. 5 displays the distribution of the size of the best subset of features chosen. The average size of this feature subset across all iterations was 5.56 features.

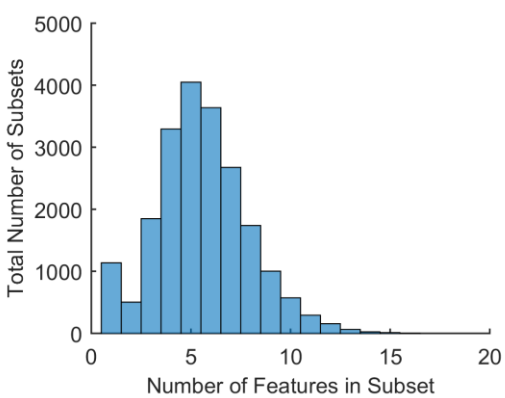

Fig. 5. Bar graph representation of the distribution, across 21,000 iterations, of subset sizes for the best feature subset and classifier combinations found during forward feature selection.

In the second analysis, after discretizing each feature (by assigning it an integer value corresponding to its quintile of the empirical distribution as in [39]), we used three different mutual information-based algorithms to produce an importance ordering of features: 1) Conditional Mutual Information (CMIM) [40]; 2) Double-Input Symmetrical Relevance (DISR) [41]; and 3) Minimal Redundancy, Maximal Relevance with differencing (MRMR) [42]. A detailed discussion of the criteria used by each of these algorithms to produce an ordered selection of features can be found in [41]. Table III shows the top five features as ordered by each algorithm. It should be noted that the ranks in Table III, similar to those in Table II, do not signify the predictive value of each feature considered individually; rather, they reflect an implicit nesting, each being contingent upon a specific set of higher-ranking features selected previously by the algorithm.

Table III. Feature Ordering Statistics

CMIM		DISR		MRMR	
Feature Name	Avg. Rank	Feature Name	Avg. Rank	Feature Name	Avg. Rank
STEF	2.75	LTS	2.51	LTS	3.17
MTEZ	11.99	SDWD4	6.37	PTS	10.22
ZCT	12.38	EWD4	7.08	SDWD4	14.84
KTZ	12.63	VWD4	8.11	HRR	15.32
ZCTZ	18.09	MTE	9.21	MTE	21.17

CMIM: Conditional Mutual Information; DISR: Double-Input Symmetrical Relevance; MRMR: Minimal Redundancy, Maximal Relevance. EWD4: Energy, Wavelet Decomposition Coefficients, Daubechies 4, Gamma; HRR: Hurst Exponent, Rescaled Range; KTZ: Kurtosis, Teager Energy Signal, Z-Score; MTE: Mean, Teager Energy Signal; MTEZ: Mean, Teager Energy Signal, Z-Score; ZCT: Number of Zero-Crossings, Time Series; ZCTZ: Number of Zero-Crossings, Time Series, Z-Score; STEF: Skewness, Teager Energy Signal, Frequency Modulated Component; SDWD4: Standard Deviation, Wavelet Decomposition Coefficients, Daubechies 4, Gamma; VTEFZ: Variance, VWD4: Variance, Wavelet Decomposition Coefficients, Daubechies 4, Gamma. See Table II for remainder of feature name abbreviations.

Results from cross-validated accuracy-based forward feature selection (Table II) and mutual information-based feature ordering (Table III) taken together indicate prominent roles for Line-Length, Number of Peaks, Teager Energy, and Wavelet-related features. Of note, all of these features are sensitive to interactions between multiple parameters (e.g., amplitude, frequency, phase) of the components of the EEG signal. This is in contrast to features such as the power in selected frequency bands, for example those used in [13], which depend only on the relative magnitudes of the components in the bands.

D. Combined Channel Analysis

We hypothesized that combining the information from all four channels in the input feature vector would improve classifier performance. To test this, all 168 features from each of the four channels were concatenated to form a single 672-element feature vector before using the nested 10-fold cross-validated forward feature selection procedure described in Section IIIB.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2016.2628702, IEEE Transactions on Cognitive and Developmental Systems

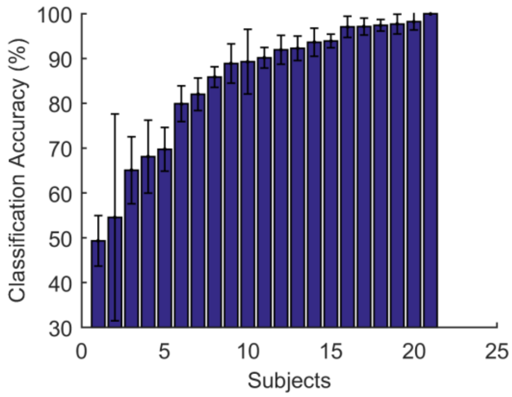

Fig. 6. Each bar represents classification accuracy for the best feature subset-classifier combination for a given subject when features from all channels were considered together in subset selection. Subjects are ordered from lowest to highest accuracy. Error bars are ± 1 standard error of the mean performance over ten cross-validation folds.

With the inclusion of features from all channels in the forward selection process, classification performance was either maintained or improved for 81.0% (17 of 21) subjects. Accuracy exceeded 75% for 16 of 21 subjects and was above 90% in 11 of 21 subjects, with 100% accuracy in one subject, as shown in fig 6. Similar to the individual-channel case, PCA was inferior to feature selection as a method of dimensionality reduction for the combined-channel case (Wilcoxon Signed Rank test, p = 0.004). Fig. 7 summarizes the results from four phases of algorithm testing: 1) three EEG power-only features with 10 classifiers as in [13]; 2) augmented set of 168 total features with 13 classifiers; 3) forward feature selection added; and 4) features combined across channels with forward feature selection. The highest performance was generally achieved using feature selection with an expanded set of features combined across channels (phase 4), though each subject maximized performance with a different combination of features and classifiers.

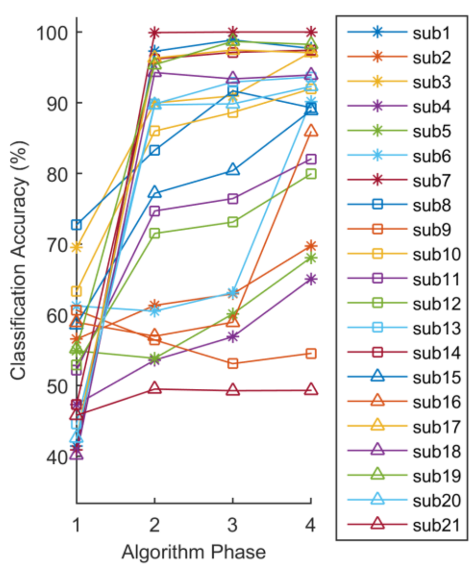

Fig. 7. Line plots showing the general improvement, across the four phases of testing described in the text, of cognitive workload classification performance. Each point shows best classification performance for a given subject. Phase 1 are results computed using the method in [13].

IV. Task Performance Assessment

We were interested in the existence and strength of any relationship between cognitive workload and task success. For example, is increasing cognitive workload always accompanied by decreasing task performance? To address this question, behavioral performance data sampled at 2 Hz were analyzed for each subject, including measurements of airspeed, heading, altitude, vertical airspeed, and angle of bank. An index characterizing behavioral performance in each flight task was developed in consultation with professional pilots and flight instructors. The first step was to set tolerances for each performance target (e.g., “must stay within ±200 ft of the assigned altitude of 4000 ft.”; “must stay within, ±10 kn of assigned airspeed of 180 kn,” etc.; See Appendix C) in order to establish a standard for acceptable performance. A curve representing the deviation from the assigned target was then generated for each performance target, bounded above and below by these acceptable performance tolerances. For each bounded curve, using one-minute windows, the ratio between the curve area corresponding to acceptable performance and the total bounded curve area was computed. Outliers were isolated and removed using a cutoff of three standard deviations from the mean over all one-minute windows, and the remaining data points were averaged to generate a score for each performance target.

A composite index representing overall task performance (e.g., collectively accounting for altitude, heading, and bank angle performance) was then computed as a normalized weighted average of the performance target scores relevant to each task (*Easy*, *Medium*, *Hard*). The composite index ranged continuously from 0 (poor task performance) to 1 (good task performance). We found a significant negative relationship between cognitive workload (task complexity level) and this composite task performance index: the Spearman’s rank correlation coefficient, computed on max-normalized data within subjects and without repeated measures to avoid violating independence assumptions, was -0.76 (n = 21; $p < 0.001$, permutation test).

Relatedly, we hypothesized that subjects who did not display a consistent decrease in behavioral performance as task complexity increased might not be sufficiently cognitively loaded and would thus not show strong EEG changes as task complexity varied. Classifier performance would therefore be expected to suffer in these subjects. To test this hypothesis, we partitioned subjects into those displaying and those not displaying a monotonic decrease in the composite performance index as task complexity increased.

Fifteen of 21 subjects (71.4%) showed such a performance decrement as task complexity increased. Fig. 8 shows classification results when this binary behavioral performance decrement variable is used to partition subjects. Notably, both groups exhibited classification performance substantially better than chance. Subjects exhibiting a decrement in behavioral performance had higher classification rates as expected, though the trend was not statistically signficant (Wilcoxon Rank Sum, one-tailed test, p = 0.185). This result taken together with the significant but imperfect negative correlation between task complexity and behavioral performance suggests that cognitive workload can increase in way that is detectable by EEG

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2016.2628702, IEEE Transactions on Cognitive and Developmental Systems

monitoring but lacking a measureable impact on behavioral performance. Such a scenario is of practical importance because it could reflect – in the absence of any overt behavioral clues – an individual's compromised ability to mount a successful response to new, emergent tasks that arise while attending to a primary task.

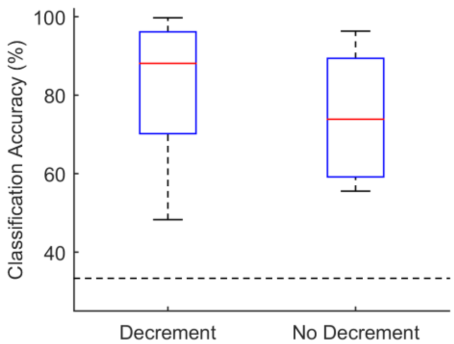

Fig. 8. Box plots showing the distribution of classification accuracies for subjects with behavioral performance decrements (n = 15) and without behavioral performance decrements (n = 6) as flight task complexity increased. The Wilcoxon Rank Sum test fails to reject the null hypothesis that the two distributions are identical (p = 0.185), though there is a trend toward higher classification accuracies in the group showing behavioral performance decrements. The dotted line represents chance performance using a null model of equiprobable classes.

V. Conclusion

The purpose of this research was to train classifiers to predict different levels of task complexity, and by extension, cognitive workload, using features extracted from background EEG collected from gel-free sensors during a flight simulator experiment. The best subsets of features for each subject yielded a median classification accuracy of 90.17%, with perfect accuracy in one subject and greater than 75% in 16 of 21 subjects. The best-performing combination of feature subset and classifier varied across subjects, suggesting that optimal performance is achieved with individualized classifiers. We did not investigate how performance would change if subject data were pooled and a single best classifier learned, though the inter-subject variability we observed suggests it would degrade. Such a "universal" cognitive workload classifier, if feasible, would have the practical advantage of eliminating a subject-specific algorithm training phase, as it could rely on data pre-recorded from a representative population of subjects.

Of note, classifiers in this study were configured using typical but untuned parameters. Optimizing parameters could yield higher classification accuracies than already achieved. Generally, the two best performing features were time-series line length and number of peaks, and the best performing classifier was LDA. Linear Discriminant Analysis often works well in practice due to its low variance, and particularly when the optimal class boundaries are linear and the class-conditional probability density functions are well described by multivariate Gaussians with a common covariance matrix, per the model's assumptions [43]. In this work, ensemble methods such as Boosted Trees and Random Forests, which are capable of learning highly nonlinear class boundaries, were often competitive with LDA, but given the relative model simplicity and computational speed of LDA we find no compelling reason to adopt a nonlinear classifier in this specific application.

No single channel proved consistently better than another for classification. However, combining all channels into a single feature vector, followed by subset selection, maintained or improved performance for 81.0% of subjects. Practical online implementation of a classifier for cognitive workload predictions may be feasible given that we used EEG measurements in 1-second historical windows, but the impacts of classifier training and testing time, as well as feature computation time, need to be further studied. In addition, achieving similar results in field-deployment will likely depend on demonstrating the robustness of classification methods to common scalp EEG artifacts excluded in this study by visual inspection, or on reliable methods for automatically rejecting them [35], [44]. We hypothesize that incorporating non-EEG features, such as eye-tracking information, skin conductance, and heart rate variability may further improve classification performance. That non-EEG measures such as heart rate variability and behavioral performance have been shown to augment the predictive power of ERP and other EEG-based classifiers supports this hypothesis [14].

We anticipate that the results of this study will guide the future implementation of cognitive workload monitoring in aviation environments. Different from prior work in this area, the present study simultaneously: 1) uses data from only a small number (four) of EEG sensors; 2) uses dry (gel-free) electrodes; 3) requires only non-evoked background EEG data; 4) analyses EEG segments of durations amenable to real-time processing; and 5) successfully classifies more than two levels of cognitive workload. We believe that all five of these aspects are critical steps toward realizing a practical cognitive workload monitoring device for pilots in the cockpit: one that is convenient to apply and comfortable to wear; computationally efficient; non-disruptive to normal operating procedures; and can predict cognitive workload levels with good resolution and accuracy. Given the high classification accuracies we achieved under these practical constraints, we believe the methods presented here will be useful in developing optimal equipment and training techniques to minimize cognitive workload in the cockpit, and to monitor the cognitive state of pilots over the course of a flight.

VI. Appendix

A. Classifier Details and Abbreviations

1. 1-Nearest Neighbor - Euclidean distance (KNN1)
2. k-Nearest Neighbor - Euclidean distance; $k \in \{3, 5, \ldots, 51\}$, chosen based on cross-validation (KNNC)
3. Linear Discriminant Analysis (LDA)
4. Quadratic Discriminant Analysis (QDA)
5. Naïve Bayes - Normal distribution assumed (NB)
6. Decision Tree (Unpruned) - Split criterion: Gini's Diversity Index (DTNP)
7. Decision Tree (Pruned) - Split Criterion: Gini's Diversity Index; Prune Criterion: Test set error (DTP)
8. Support Vector Machine - Kernel: Radial Basis Function; One vs. all (SVMR)
9. Support Vector Machine - Kernel: Linear; One vs. all (SVML)
10. Support Vector Machine - Kernel: Polynomial, order 2 or 3 chosen using cross-validation; One vs. all (SVMP)

11. Random Forest (RF)
12. Boosted Decision Tree (AdaBoost.M2 algorithm) (BT)
13. Subspace Linear Discriminant Analysis (SDA)

B. Feature Descriptions

This section gives mathematical expressions for the two features most frequently chosen by forward stepwise selection, followed by a comprehensive listing, in compact form, of all 168 computational features used in this work. The former, as indicated in Table II are time-series line length and number of peaks. Line length, L, is the sum of the absolute values of the differences between successive samples in a given N-sample EEG segment, $x[n]$:

$$L = \sum_{n=0}^{N-2} |x[n+1] - x[n]| \tag{1}$$

The line length feature tends to be sensitive to signal amplitude and frequency variation, and was originally developed for use in detecting the onset of epileptic seizures [45]. The number of peaks, P, counts the number of times the first-order hold, $x(t)$, of a given N-sample EEG segment changes from positive to negative slope:

$$P = \frac{1}{2}\int_{0}^{(N-1)T} \left\{\frac{d[sgn(x'(t))]}{dt}\right\}^{-} dt \tag{2}$$

with T the sampling period and $\{\cdot\}^{-}$ the negative-part operator.

Below we provide a compact listing of all features used in this work followed by a glossary of terms that helps interpret the descriptions and gives references to the literature.

1. *Average power* with/without Z-scoring (delta, theta, alpha, beta, and gamma bands): 10 features
2. *Power spectral density* with/without Z-scoring (mean, standard deviation, variance, skewness, kurtosis): 10 features
3. *Number of zero-crossings* (time-series) with/without Z-scoring: 2 features
4. *Number of peaks* (time-series) with/without Z-scoring: 2 features
5. *Teager energy signal & Teager energy signal frequency modulated component* with/without Z-scoring (mean, standard deviation, variance, skewness, kurtosis): 20 features
6. *Line length* (time-series) with/without Z-scoring: 2 features
7. *Hurst exponent* with/without Z-scoring (discrete second-order derivative, wavelet-based adaptation, rescaled range): 6 features
8. *Wavelet decomposition-based features.* Coefficients (Daubechies 4 wavelet) for delta, theta, alpha, beta, gamma bands with/without Z-scoring (maximum, minimum, mean, standard deviation, variance, skewness, kurtosis, energy, Recoursing Energy Efficiency (REE), Logarithmic REE, Absolute Logarithmic REE): 110 features
9. *Correlation dimension* using Grassberger-Procaccia Algorithm with/without Z-scoring: 2 features
10. *Entropy* (time-series) with/without Z-scoring: 2 features
11. *Optimum minimum time-delay from phase-space reconstruction* (time-series) with/without Z-scoring: 2 features

<u>*Absolute Logarithmic Recoursing Energy Efficiency:*</u> Absolute value of the Logarithmic Recoursing Energy Efficiency [28]

<u>*Alpha band:*</u> Frequencies in the range 8 – 13 Hz

<u>*Beta band:*</u> Frequencies in the range 13 – 30 Hz

<u>*Correlation dimension:*</u> The dimension of an attractor in the phase space of a dynamical system, estimated from a sample of points on the attractor [46]

<u>*Delta band*</u>: Frequencies in the range 1 – 4 Hz.

<u>*Entropy:*</u> A measure of the uncertainty in a random variable. For a discrete random variable *X* with alphabet *X'* and probability mass function $p_X(x)$, the Entropy, H(*X*), is defined as: $H(X) = -\sum_{x \in X'} p_X(x) log_2 p_X(x)$ [47], [48]

<u>*Frequency modulated component of Teager energy signal:*</u> Nonlinear combinations of instantaneous Teager energy signal outputs to extract the frequency-modulated component [49][50] of the signal; originally used in speech processing

<u>*Gamma band:*</u> Frequencies in the range 30 – 40 Hz.

<u>*Grassberger-Procaccia algorithm*</u>: Method for estimating the correlation dimension [46]

<u>*Hurst exponent*</u>: Measure of the smoothness of the fractal dimension of a time-series. The fractal dimension quantifies the complexity of a signal [51]

<u>*Kurtosis:*</u> The fourth standardized moment of a random variable; a measure of the "tailedness" of the distribution

<u>*Logarithmic Recoursing Energy Efficiency:*</u> Base-10 logarithm of the REE [52]

<u>*Optimum minimum time-delay:*</u> Optimal time delay to use for phase-space reconstruction [53]

<u>*Phase-space reconstruction:*</u> Using time-delayed versions of a signal to generate coordinates for a phase-space representation of a system in order to study its dynamics [53]

<u>*Power spectral density:*</u> The Fourier transform (when it exists) of the autocorrelation function of a stationary random process; quantifies how the power of a signal is distributed across a range of frequencies [54]

<u>*Recoursing Energy Efficiency:*</u> Measures the percentage of energy calculated from the wavelet decomposition coefficients of a signal within a sub-band [52]

<u>*Skewness:*</u> The third standardized moment of a of a random variable; measures the asymmetry of the distribution about the mean [55]

<u>*Teager Energy:*</u> A measure developed to approximate the energy required to generate a signal by mechanical analogy [56].

<u>*Theta band:*</u> Frequencies in the range 4 – 8 Hz

<u>*Wavelet decomposition coefficients:*</u> The coefficients calculated from the decomposition of signal into a series of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2016.2628702, IEEE Transactions on Cognitive and Developmental Systems

basis functions called wavelets [57]; these coefficients represent dilations, contractions, and shifts of a prototype signal (mother wavelet)

Z-scoring: Standardizing a data set by subtracting the sample mean from each point and dividing by the sample standard deviation

C. Acceptable Performance Criteria

This section describes the criteria used to define acceptable performance on each of the three flight tasks.

1. *Easy*: no more than: ±200 ft from assigned altitude, ±10 kn from assigned airspeed, ±5º from assigned heading, and ±5 º from assigned bank angle.
2. *Medium*: no more than: ±200 ft from assigned altitude, ±10 kn from assigned airspeed, ±5º from assigned heading, ±5º from assigned bank angle, and ±500 fpm from assigned ascent and descent rates, respectively.
3. *Hard*: no more than: ±200 ft from assigned altitude, ±10 kn of assigned airspeed, ±5 º from assigned heading, ±5º from assigned bank angle, and ±500 fpm from assigned ascent and descent rates, respectively.

VII. Acknowledgment

We thank Dr. Jessica Mohler, CAPT Mike Prevost, CAPT Kenneth Ham, and CDR Hartley Postlethwaite from the United States Naval Academy in Annapolis, MD for their research coordination efforts and aviation advice.

VIII. References

[1] B. H. Kantowitz, "Mental Workload," in *Advances in Psychology: Human Factors Psychology*, 1987, pp. 81–121.

[2] G. F. Hankins, T.C., Wilson, "A comparison of heart rate, eye activity, EEG and subjective measures of pilot mental workload during flight," *Aviat. Space. Environ. Med.*, 1998.

[3] P. Besson, C. Bourdin, L. Bringoux, E. Dousset, C. Maïano, T. Marqueste, D. R. Mestre, S. Gaetan, J. Baudry, and J. Vercher, "Effectiveness of Physiological and Psychological Features to Estimate Helicopter Pilots' Workload : A Bayesian Network Approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 4, pp. 1872–1881, 2013.

[4] W. Li, F. Chiu, K. Wu, and B. District, "The Evaluation of Pilots Performance and Mental Workload by Eye Movement," in *Proceedings of the 30th European Association for Aviation Psychology*, 2012.

[5] M. A. Hogervorst, A.-M. Brouwer, and J. B. F. van Erp, "Combining and comparing EEG, peripheral physiology and eye-related measures for the assessment of mental workload.," *Front. Neurosci.*, vol. 8, p. 322, Jan. 2014.

[6] T. K. Calibo, J. A. Blanco, and S. L. Firebaugh, "Cognitive stress recognition," in *2013 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, 2013, pp. 1471–1475.

[7] A. Uetake and A. Murata, "Assessment of mental fatigue during VDT task using event-related potential (P300)," in *Proceedings of the 2000 IEEE International Workshop on Robot and Human Interactive Communication*, 2000, pp. 235–240.

[8] M. A. Hogervorst, A. Brouwer, and J. B. F. Van Erp, "Combining and comparing EEG , peripheral physiology and eye-related measures for the assessment of mental workload," *Front. Neurosci.*, vol. 8, no. October, pp. 1–14, 2014.

[9] T. J. Sullivan, S. R. Deiss, T. P. Jung, and G. Cauwenberghs, "A brain-machine interface using dry-contact, low-noise EEG sensors," *Proc. - IEEE Int. Symp. Circuits Syst.*, pp. 1986–1989, 2008.

[10] R. J. Gentili, J. C. Rietschel, K. J. Jaquess, L. C. Lo, M. Prevost, M. W. Miller, J. M. Mohler, H. Oh, Y. Y. ing Tan, and B. D. Hatfield, "Brain biomarkers based assessment of cognitive workload in pilots under various task demands," *Conf. Proc. ... Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Conf.*, vol. 2014, pp. 5860–5863, 2014.

[11] W. S. Pritchard, "Psychophysiology of P300," *Psychol. Bull.*, vol. 89, no. 3, pp. 506–540, 1981.

[12] M. W. Miller, J. C. Rietschel, C. G. McDonald, and B. D. Hatfield, "A novel approach to the physiological measurement of mental workload.," *Int. J. Psychophysiol.*, vol. 80, no. 1, pp. 75–8, Apr. 2011.

[13] M. K. Johnson, J. A. Blanco, R. J. Gentili, K. J. Jaquess, H. Oh, and D. Hatfield, "Probe-Independent EEG Assessment of Mental Workload in Pilots," *7th Annu. Int. IEEE EMBS Conf. Neural Eng.*, pp. 22–24, 2015.

[14] H. Oh, B. D. Hatfield, K. J. Jaquess, L. Lo, Y. Ying, M. C. Prevost, J. M. Mohler, H. Postlethwaite, C. Rietschel, M. W. Miller, J. A. Blanco, S. Chen, and R. J. Gentili, "A Composite Cognitive Workload Assessment System in Pilots Under Various Task Demands using Ensemble Learning," in *Foundations of Augmented Cognition, Lecture Notes in Computer Science*, 2015, pp. 91–100.

[15] G. Borghini, L. Astolfi, G. Vecchiato, D. Mattia, and F. Babiloni, "Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness.," *Neurosci. Biobehav. Rev.*, pp. 1–18, Oct. 2012.

[16] C. Dussault, J. C. Jouanin, and C. Y. Guezennec, "EEG and ECG changes during selected flight sequences," *Aviat. Sp. Environ. Med.*, vol. 75, no. 10, pp. 889–897, 2004.

[17] C. Dussault, J. C. Jouanin, M. Philippe, and C. Y. Guezennec, "EEG and ECG changes during simulator operation reflect mental workload and vigilance," *Aviat. Sp. Environ. Med.*, vol. 76, no. 4, pp. 344–351, 2005.

[18] Z. J. Koles and P. Flor-Henry, "Mental activity and the EEG: task and workload related effects," *Med. Biol. Eng. Comput.*, vol. 19, no. 2, pp. 185–194, 1981.

[19] S. H. Fairclough, L. Venables, and A. Tattersall, "The influence of task demand and learning on the psychophysiological response," *Int. J. Psychophysiol.*, vol. 56, no. 2, pp. 171–184, 2005.

[20] C. Berka, D. J. Levendowski, M. N. Lumicao, A. Yau, G. Davis, V. T. Zivkovic, R. E. Olmstead, P. D. Tremoulet, and P. L. Craven, "EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks.," *Aviat. Space. Environ. Med.*, vol. 78, no. 5 Suppl, pp. B231–B244, 2007.

[21] A. Sassaroli, F. Zheng, L. M. Hirshfield, A. Girouard, E. T. Solovey, R. J. K. Jacob, and S. Fantini, "Discrimination of mental workload levels in human subjects with functional near-infrared spectroscopy," *J. Innov. Opt. Health Sci.*, vol. 1, no. 02, pp. 227–237, 2008.

[22] G. F. Wilson and C. A. Russell, "Real-Time Assessment of Mental Workloa Using Psychophysiological Measures and Artificial Neural Networks," vol. 298, no. 074, 2004.

[23] K. R. Muller, M. Tangermann, G. Dornhege, M. Krauledat, G. Curio, and B. Blankertz, "Machine learning for real-time single-trial EEG-analysis: From brain-computer interfacing to mental state monitoring," *J. Neurosci. Methods*, vol. 167, no. 1, pp. 82–90, 2008.

[24] J. B. Noel, K. W. Bauer, and J. W. Lanning, "Improving pilot mental workload classification through feature exploitation and combination: a feasibility study," *Comput. Oper. Res.*, vol. 32, no. 10, pp. 2713–2730, Oct. 2005.

[25] T. McMahan, T. D. Parsons, and I. Parberry, "Modality Specific Assessment of Video Game Player's Cognitive Workload Using Off-the-Shelf Electroencephalographic Technologies," 2014.

[26] J. C. Rietschel, M. W. Miller, R. J. Gentili, R. N. Goodman, C. G. McDonald, and B. D. Hatfield, "Cerebral-cortical networking and activation increase as a function of cognitive-motor task difficulty," *Biol. Psychol.*, vol. 90, no. 2, pp. 127–33, May 2012.

[27] M. M. Shaker, "EEG waves classifier using wavelet transform and Fourier transform," *World Acad. Sci. Eng. Technol.*, vol. 3, pp. 723–728, 2007.

[28] M. Murugappan, "Classification of human emotion from EEG using discrete wavelet transform," *J. Biomed. Sci. Eng.*, vol. 03, no. 04, pp. 390–396, 2010.

[29] K. Brookhuis and D. de Waard, "Monitoring drivers' mental workload in driving simulators using physiological measures," *Accid. Anal. Prev.*, no. 42, pp. 898–903, 2010.

[30] H. Petkar, R. Yadav, T. A. Nguyen, Y. Zeng, and S. Dande, "A Pilot Study to Assess Designer's Mental Stress using Eye Caze System

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2016.2628702, IEEE Transactions on Cognitive and Developmental Systems

and Electroencephalogram," *Proc. ASME 2009 Int. Des. Eng. Tech. Conf. Comput. Inf. Eng. Conf.*, pp. 1–11, 2009.

[31] D. Dasari, G. Shou, and L. Ding, "Investigation of independent components based EEG metrics for mental fatigue in simulated ATC task," in *6th Annual International Conference on Neural Engineering*, 2013, pp. 1331–1334.

[32] E. B. J. Coffey, a.-M. Brouwer, and J. B. F. van Erp, "Measuring workload using a combination of electroencephalography and near infrared spectroscopy," *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, vol. 56, no. 1, pp. 1822–1826, Oct. 2012.

[33] P. Zarjam, J. Epps, and N. H. Lovell, "Beyond Subjective Self-Rating : EEG Signal Classification of Cognitive Workload," *IEEE Trans. Auton. Ment. Dev.*, vol. 7, no. 4, pp. 301–310, 2015.

[34] V. Jurcak, D. Tsuzuki, and I. Dan, "10/20, 10/10, and 10/5 Systems Revisited: Their Validity As Relative Head-Surface-Based Positioning Systems.," *Neuroimage*, vol. 34, no. 4, pp. 1600–11, Feb. 2007.

[35] L. Sornmo and P. Laguna, *Bioelectrical Signal Processing in Cardiac and Neural Applications*. Burlington, MA: Elsevier Academic Press, 2005.

[36] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*. New York, NY: Springer Science+Business Media, LLC, 2013.

[37] S. Varma and R. Simon, "Bias in error estimation when using cross-validation for model selection.," *BMC Bioinformatics*, vol. 7, p. 91, 2006.

[38] K. Pearson, "On Lines and Planes of Closest Fit to Systems of Points in Space," *Philos. Mag.*, vol. 2, no. 11, pp. 559–572, 1901.

[39] N. X. Vinh, J. Chan, and J. Bailey, "Reconsidering Mutual Information Based Feature Selection : A Statistical Significance View," *Twenty-Eighth AAAI Conf. Artif. Intell.*, no. 1, 2014.

[40] F. Fleuret, "Fast Binary Feature Selection with Conditional Mutual Information," *J. Mach. Learn. Res.*, vol. 5, pp. 1531–1555, 2004.

[41] P. E. Meyer and G. Bontempi, "On the use of variable complementarity for feature selection in cancer classification," *Lect. Notes Comput. Sci.*, vol. 3907, pp. 91–102, 2006.

[42] H. C. Peng, F. H. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, 2005.

[43] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. New York, NY: Springer-Verlag, 2001.

[44] J. A. Urigüen and B. Garcia-Zapirain, "EEG artifact removal-state-of-the-art and guidelines.," *J. Neural Eng.*, vol. 12, no. 3, p. 31001, 2015.

[45] D. Olsen, R. P. Lesser, J. C. Harris, W. R. S. Webber, and J. A. Cristion, "Automatic detection of seizures using electroencephalographic signals," USPTO Pat. No. 5311876, 1992.

[46] J. Theiler, "Efficient algorithm for estimating the correlation dimension from a set of discrete points," *Physical Review A*, vol. 36. pp. 4456–4462, 1987.

[47] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis*. 2003.

[48] N. Kannathal, S. K. Puthusserypady, and L. Choo Min, "Complex dynamics of epileptic EEG.," *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, vol. 1, pp. 604–7, Jan. 2004.

[49] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy Separation in Signal Modulations with Application to Speech Analysis," *IEEE Trans. Signal Process.*, vol. 41, no. 10, pp. 3024 – 3051, 1993.

[50] G. Zhou, J. Hansen, and J. Kaiser, "Nonlinear feature based classification of speech under stress," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 3, pp. 201–216, 2001.

[51] M. Kale and F. Butar, "Fractal Analysis of Time Series and Distribution Properties of Hurst Exponent," *J. Math. Sci. Math. Educ.*, vol. 5, no. 1, pp. 8–19.

[52] M. Murugappan, M. Rizon, R. Nagarajan, S. Yaacob, D. Hazry, and I. Zunaidi, "Time-Frequency Analysis of EEG Signals for Human Emotion Detection," *4th Kuala Lumpur Int. Conf. Biomed. Eng. 2008 IFMBE*, no. c, pp. 262–265, 2008.

[53] M. B. Kennel, R. Brown, and H. D. I. Abarbanel, "Determining embedding dimension for phase-space reconstruction using a geometrical construction," *Physical Review A*, vol. 45. pp. 3403–3411, 1992.

[54] P. Welch, "The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE Trans. audio Electroacoust.*, 1967.

[55] Y. Maruyama, "Measures of Multivariate Skewness and Kurtosis with Principal Components," *Japanese J. Appl. Stat.*, vol. 36, no. 3, pp. 139–145, 2007.

[56] J. F. Kaiser, "On a Simple Algorithm to Calculate the Energy of a Signal," in *Acoustics, Speech, and Signal Processing, 1988*, 1990, pp. 381–384.

[57] M. Akin, "Comparison of wavelet transform and FFT methods in the analysis of EEG signals.," *J. Med. Syst.*, vol. 26, no. 3, pp. 241–7, Jun. 2002.