AFRL-RH-WP-TR-2013-0007

THE EFFECTS OF DAY-TO-DAY VARIABILITY OF PHYSIOLOGICAL DATA ON OPERATOR FUNCTIONAL STATE CLASSIFICATION

James C. Christensen, Justin R. Estepp, Glenn F. Wilson,
Christopher A. Russell, & Krystal M. Thomas
Warfighter Interface Division

March 2013

FINAL REPORT

AIR FORCE RESEARCH LABORATORY
711 HUMAN PERFORMANCE WING,
HUMAN EFFECTIVENESS DIRECTORATE,
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433
AIR FORCE MATERIEL COMMAND
UNITED STATES AIR FORCE

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report was cleared for public release by the 88th Air Base Wing Public Affairs Office and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RH-WP-TR-2013-0007 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

THOMAS.KRYSTAL.M.1275004304
Digitally signed by THOMAS.KRYSTAL.M.1275004304
DN: c=US, o=U.S. Government, ou=DoD, ou=PKI, ou=USAF, cn=THOMAS.KRYSTAL.M.1275004304
Date: 2013.03.26 16:16:03 -04'00'

Krystal M. Thomas
Work Unit Manager
Collaborative Interfaces Branch

RUSSELL.WILLIAM.E.JR.1007068057
Digitally signed by RUSSELL.WILLIAM.E.JR.1007068057
DN: c=US, o=U.S. Government, ou=DoD, ou=PKI, ou=USAF, cn=RUSSELL.WILLIAM.E.JR.1007068057
Date: 2013.03.27 07:58:13 -04'00'

William E. Russell
Chief, Collaborative Interfaces Branch
Warfighter Interface Division

FREEMAN.GEORGE.R.1019635976
Digitally signed by FREEMAN.GEORGE.R.1019635976
DN: c=US, o=U.S. Government, ou=DoD, ou=PKI, ou=USAF, cn=FREEMAN.GEORGE.R.1019635976
Date: 2013.03.27 10:25:53 -04'00'

G. Richard Freeman
Warfighter Interface Division
Human Effectiveness Directorate
711 Human Performance Wing

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
01-03-13	Final	14 October 2009 – 28 December 2012

4. TITLE AND SUBTITLE

THE EFFECTS OF DAY-TO-DAY VARIABILITY OF PHYSIOLOGICAL DATA ON OPERATOR FUNCTIONAL STATE CLASSIFICATION

5a. CONTRACT NUMBER
IN-HOUSE

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
62202F

6. AUTHOR(S)

James C. Christensen, Justin R. Estepp, Glenn F. Wilson, Christopher A. Russell, & Krystal M. Thomas

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER
(H02W) 53290802

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Applied Adaptive Aiding Section, 711th Human Performance Wing
Air Force Research Laboratory
2510 Fifth Street, B840, W200
Wright-Patterson AFB, OH 45433-7951

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Materiel Command
Air Force Research Laboratory
711 Human Performance Wing
Human Effectiveness Directorate
Warfighter Interface Division
Applied Neuroscience Branch
Applied Adaptive Aiding Section
Wright-Patterson Air Force Base, OH 45433

10. SPONSORING/MONITORING AGENCY ACRONYM(S)
711HPW/RHCPA

11. SPONSORING/MONITORING AGENCY REPORT NUMBER(S)
AFRL-RH-WP-TR-2013-0007

12. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution A: Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES
88 ABW Cleared 3/26/2013; 88ABW-2013-1480. Report contains color.

14. ABSTRACT

The application of pattern classification techniques to physiological data has undergone rapid expansion. Tasks as varied as the diagnosis of disease from magnetic resonance images, brain-computer interfaces for the disabled, and the decoding of brain functioning based on electrical activity have been accomplished quite successfully with pattern classification. These classifiers have been further applied in complex cognitive tasks to improve performance, in one example as an input to adaptive automation. In order to produce generalizable results and facilitate the development of practical systems, these techniques should be stable across repeated sessions. This paper describes the application of three popular pattern classification techniques to EEG data obtained from asymptotically trained subjects performing a complex multitask across five days in one month. All three classifiers performed well above chance levels. The performance of all three was significantly negatively impacted by classifying across days; however two modifications are presented that substantially reduce misclassifications. The results demonstrate that with proper methods, pattern classification is stable enough across days and weeks to be a valid, useful approach.

15. SUBJECT TERMS
EEG, pattern classification, interday variability, workload, day-to-day variability

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT:	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON (Monitor)
a. REPORT	b. ABSTRACT	c. THIS PAGE			Krystal Thomas
Unclassified	Unclassified	Unclassified	SAR	25	19b. TELEPHONE NUMBER *(Include Area Code)*

This page intentionally left blank.

TABLE OF CONTENTS

Section **Page**

LIST OF FIGURES AND TABLES

Page

ACKNOWLEDGEMENTS

The authors would like to thank Samantha Klosterman, Jason Monnin, and Krystal Thomas for their assistance in data analysis and the preparation of this manuscript.

1.0 SUMMARY

This report describes work performed under the "Adaptive Aiding for Warfighter Operations" work unit. As originally conceived, this work was intended to advance the state of the art in neurophysiological triggering of adaptive aiding. Early results indicated that the most significant roadblock to this triggering scheme was poor stability of existing techniques when applied over longer time periods, such as days or weeks. Consequently, work focused on addressing this issue. Over the course of this effort, two techniques for enhancing the stability of workload monitoring via pattern classification of neurophysiology were identified and demonstrated to be effective. The first technique is to collect baselines over multiple days. The second is to collect a small (5 minutes) amount of baseline data for each new day that you wish to run the system. When used together, stability over days and weeks rises to the same level as stability within a few hours, and is likely adequate for future applications.

2.0 INTRODUCTION

The application of pattern classification to physiological data has become increasingly popular. This includes a wide range of areas such as brain-computer interfaces (BCI, reviewed in Birbaumer, 2006), neurology (Blanco, et al., 2010), psychiatry (Coburn, et al., 2006) and multi-voxel pattern analysis of fMRI data (e.g. Kamitami & Tong, 2005). This approach has been remarkably successful in classifying mental workload in complex tasks (Berka, et al., 2004; Freeman, Mikulka, Prinzel & Scerbo, 1999; Gevins, et al., 1998; Wilson & Fisher, 1991; Wilson & Russell, 2003a; 2003b). Further, this information has been used to modify an operator's task via adaptive aiding with the goal of enhancing overall performance in demanding cognitive workload situations (Freeman, Mikulka, Prinzel & Scerbo, 1999; Wilson & Russell, 2007). In this last line of research, the focus on more realistic, complex tasks and the possibility of improved performance have rendered it very much in line with the concepts of neuroergonomics (Parasuraman & Wilson, 2008). For example, Wilson and Russell (2007) utilized a complex uninhabited aerial vehicle simulation to show that physiologically driven adaptive aiding could improve overall performance. Operator physiology was monitored and used to discriminate between task demand levels. The task was presented at two levels of difficulty and electroencephalographic (EEG), electrooculographic (EOG), and cardiac data were recorded while the operators performed the task. These data were used to train an artificial neural network (ANN) to recognize patterns in the physiological data that corresponded to the performance of the low and high mental demand conditions. The operators then performed the task again and adaptive aiding was provided when the classifier determined that they were experiencing the high workload situation. The aiding intervention was such that the operators were given more time to evaluate possible target stimuli. The physiologically driven adaptive aiding improved their performance by approximately 50%. This approach permitted the coupling of operator and system so that the momentary capabilities of the operator were monitored and used to determine whether or not they needed automation assistance. This provides the groundwork for systems that would be capable of monitoring operator functional state (OFS) and modifying task demands to assist the operator in times of cognitive overload. These systems should produce improved overall effectiveness and potentially reduce catastrophic errors in real-world situations.

The classification approach to mental state estimation sidesteps some of the issues associated with multiple comparisons common in high-dimensional physiological data, though it does carry with it

potential confounds such as the effects of data overfitting. Ideally, independent samples should be used for training and testing classifiers to produce robust results (Kriegeskorte, et al., 2009). Whether the goal is to build robust adaptive systems or to obtain robust results from independent samples in a classification study, it is necessary to collect data at multiple intervals separated in time. This raises the possibility that either the collection methods or the phenomena of interest are not stable across that time period: "An important and unresolved question is the extent to which classification-based decoding strategies might generalize over time, across subjects and to new situations" (Haynes & Rees, 2006). With regard to complex task performance in applied settings where adaptive aiding may be implemented, the OFS monitor must function properly every day in order to be useful. Therefore, it is necessary to determine the stability of the physiological signals over time during complex task performance. Further, since the physiological data are used as input variables for a classifier, the output of the classifiers must be evaluated to test their reliability. To date, the effects of day-to-day fluctuations in the operator's physiology have not been thoroughly assessed while operators are engaged in complex tasks.

The stability of EEG signals has been investigated using eyes open/eyes closed conditions or while operators were engaged in simple laboratory tasks. In those contexts, the EEG has been found to be fairly stable over time within each individual (Burgress & Gruzelier, 1993; McEvoy, Smith & Gevins, 2000; Pollock, Schneider, & Lyness, 1991; Salinsky, Oken, & Morehead, 1991). These previous studies relied upon spectral comparison rather than classification. In previous research examining the stability of fMRI results as a function of analysis technique and day (McGonigle et al., 2000; Smith et al., 2005), between session (and day) variance was found to be comparable to within session variance. However, reliability generally decreased with task complexity. BCI systems also recognize the deleterious effects of day-to-day variation in the EEG signals and include procedures to ameliorate these effects (Wolpaw, et al., 2002). Huang et al (2011) present a procedure based on single-trial classification of event-related potentials (ERPs) in a target-detection task. We would expect that ERP components, such as P300, that are associated with rare targets should exhibit little variability from day to day, they were nonetheless able to show that incrementally adding additional sessions to their training set produced statistically significant increases in classifier performance, with area under the ROC curves increasing from approximately .95 to .98. It is unknown to what extent this result will apply to more complex tasks that cannot be structured as an ERP design, but must instead rely on spectral features for classification.

The assessment of reliability over time has most commonly been conducted by comparing one session to another within a day. However, adaptive systems require continuous, near real-time estimates of OFS that may exhibit greater variability. Further, the reliability of ensembles of input features as assembled by the classifier may not be predictable from knowledge of the reliability of each input feature alone. ANN and linear discriminant analysis (LDA) have been used to determine OFS (Berka, et al., 2004; Wilson & Fisher, 1991; Wilson & Russell, 2003a), while kernel-based support vector machines (SVM) have been demonstrated to be effective in classifying physiological data (e.g. De Martino et al., 2007; Garrett et al, 2003; Lal et al., 2004). Poggio et al. (2004) have shown that classifiers that are stable under leave-one-out validation with stable error are optimally generalizable; consequently, an incremental SVM with leave-one-out optimization (Cauwenberghs & Poggio, 2001) would be expected to generalize well as long as test data are drawn from the same underlying class distributions. In order to reduce the possibility that the present results are unique to the method chosen, all of these methods will be used and compared to test the reliability of OFS determination methods on the scales of seconds, hours, days, and weeks.

Classification that does not generalize across days could be indicative that the classifier has become too dependent on unique or spurious differences between classes in the training set, known as overfitting. If the training data are drawn from just one day, then the classifier may key in on unstable features unique to that day. An obvious solution to the problem of overfitting is to use multiple days in the training set, which should improve generalization. This will be tested in detail. The primary purpose of this study was to assess the stability of human operator physiology, as decoded by pattern classifiers, while performing a complex task over a four week period. We chose to focus on electrophysiology, as the collection conditions may be more carefully controlled across days than fMRI and it is more amenable to operational settings. We, therefore, set out to test the consequences of gathering and classifying electrophysiological data from multiple days in the context of OFS classification.

3.0 METHODS

The Multi-Attribute Task Battery (MATB) was used to provide three levels of complex task difficulty (Comstock & Arnegard, 1992). The task is broadly representative of aircraft operation (particularly remote piloting), and can include compensatory manual tracking, visual and auditory monitoring, and a dynamic resource allocation task. For this study, the monitoring (lights, dials, and communications) and resource allocation (fuel management) tasks were presented simultaneously during all task conditions. The compensatory manual tracking task was fully automated to more closely simulate advanced remotely piloted aircraft interfaces. The demands of each task were varied so that overall, three levels of difficulty were available. Reaction times were collected from the monitoring and communication tasks and error scores were calculated from the resource allocation task. Eight adult subjects (3 male), mean age of 21.1 years, were trained on the MATB until performance parameters attained asymptote with minimal errors. This procedure helped to reduce learning effects and allowed subjects to reach a desired level of familiarity and comfort with the laboratory setting. This training took approximately 3 hours over one or two days.

During each recording session, subjects were presented a randomized sequence of low, medium and difficult task levels. Sessions consisting of five minutes each of low, medium and high cognitive load in random order were presented three times on each of five days distributed over one month. The number of days between sessions was randomized across subjects, with each participant assigned to a random order of four intervals: one day apart, two weeks apart, and two, one-week-apart intervals. Two subjects' testing sessions are depicted in Figure 1. This randomization was intended to reduce the effects of fatigue and strategic changes associated with concentrated data collection.

SUN	MON	TUE	WED	THU	FRI	SAT
Week 1	Day 1			Day 1	Day 2	
Week 2	Day 2				Day 3	
Week 3						
Week 4	Day 3	Day 4			Day 4	
Week 5		Day 5			Day 5	

Figure 1. Data collection days for two representative subjects. The intervals between days were randomized from a set of four intervals: one day apart, two weeks apart, and two, one-week-apart intervals.

Physiological data were recorded from the subjects during task performance. Nineteen channels of EEG data were recorded at sites positioned according to the International 10-20 electrode system (Jasper, 1958). Mastoids were used as reference and ground with electrode impedances measured and maintained below 5 kOhms. Horizontal and vertical EOG and electrocardiogram (ECG) were also recorded. Each EEG channel (sampled at 256 Hz) was corrected for eye movement and blinks using a post-hoc regression method. The time series EEG, horizontal and vertical EOG data were filtered using elliptical IIR filter banks with passbands consistent with the traditional EEG bands. The frequency ranges of the five bands of EEG were delta (0.5-3 Hz), theta (4-7 Hz), alpha (8-12 Hz), beta (13-30 Hz) and gamma (31-42 Hz). Additionally, two expanded gamma bands were used, 32 to 58 Hz and 63 to 100 Hz. Waveform length was also calculated for each EEG channel, in both one second and 10 second epochs (Pleydell-Pearce, Whitecross & Dickson, 2003; Shelley & Backs, 2006). The raw ECG waveform was post-processed to extract time between successive R-wave peaks. The raw VEOG waveform was used to post-process a blink rate data channel. Blinks were automatically detected using the algorithm developed by Kong & Wilson (1998); eyeblink duration and amplitude were then extracted per their suggestion of using the half-amplitude technique.

All of the features were segmented into 40-second windows with a 35-second overlap, producing a consistent sampling rate. All band power and waveform length features, when combined, formed a bank of 189 features (9 features for each of the 21 EEG/EOG channels). The four additional peripheral features - cardiac interbeat intervals, blink rate, blink amplitude and blink duration - were also used as input features, resulting in 193 total features.

In order to estimate the functional state of the operators, three classifiers were used to classify the physiological data on an individual subject level: ANN, SVM, and LDA. Equal numbers of exemplars from the low and high cognitive load conditions were used to train the classifiers, representing data from easy and difficult task conditions, respectively. Data from the medium cognitive load portion of the tasks were omitted from analysis to facilitate a binary data classification

paradigm. The same preprocessed psychophysiological data was provided to all three classifiers, split into training, test, and validation sets where appropriate. Fifty percent of the data samples associated with any given analysis were randomly selected and used for classifier training, while twenty five percent were used to test the trained classifiers' ability to identify the easy and difficult conditions. The remaining twenty five percent was held back as a validation set to control overfitting. The training and test sets included various combinations of days and sessions within a day, as detailed in the results. The data in each of the training sets were normalized separately for each feature by first dropping the highest and lowest five percent of data points to reduce the impact of outliers, and then extracting means and standard deviations. These parameters were then used to normalize both the training and validation sets to zero mean and unit standard deviation. Test sets were separately normalized to themselves using the same procedure.

The ANN was a feedforward backpropagation neural network (Widrow and Lehr, 1990; Lippmann, 1987) implemented via the MATLAB Neural Network Toolbox (MATLAB R2008a, Neural Network Toolbox Version 6.0, The Mathworks, Natick, MA). First, the network learned the input-output classification from a set of training vectors. A separate validation set was used during training in order to reduce overfitting (Wilson & Russell, 2003a; Bishop, 2006): for any given learning iteration, the weights and biases of the ANN (derived from learning on the training set) were updated only if the feed-forward error on the validation set was equal to or less than the validation error obtained in the previous iteration. Once trained, network weights were fixed and the ANN acted as a feed-forward pattern classifier. As a classifier, the network examined input data it had never seen and predicted the class of the input data as either easy or difficult.

SVMs were constructed and tested using both the kernel-based least-squares SVM (LS-SVM) formulation presented by Suykens et al. (2002) and the incremental/decremental method with leave-one-out validation from Cauwenberghs and Poggio (2001). Lacking a priori evidence in favor of any one input kernel, linear and tuned Gaussian radial basis function (GRBF) kernels were evaluated. The tuning parameters for the GRBF were determined individually via grid search optimization on each of the training sets, as implemented in the LS-SVM MATLAB toolbox. As SVM construction was not iterative with a stopping rule like the ANNs, the validation set was added to the training set.

The LDA was calculated using the implementation found in the MATLAB Statistics Toolbox (MATLAB R2008a, Statistics Toolbox Version 6.2, The Mathworks, Natick, MA). The same training and test data sets used by the ANN and SVM were used by the LDA. The test data sets were again used to determine how accurately the trained classifier could correctly identify which of these data were from low or high cognitive load conditions. As with the implementation of the SVMs, the validation set was included as part of the training data.

4.0 RESULTS

Analysis of the performance data revealed that the easy and difficult conditions produced significantly different mean responses (see Table 1). For the communication, dials and lights tasks the difficult task produced significantly longer reaction times. The difficult task also resulted in significantly greater error scores in the resource management task. While the two levels of task difficulty produced significantly different operator performance, the main effect for days was not significant in any cases. The interaction between difficulty and days was likewise not significant.

This suggests that there was no significant change in task performance across days, a critical condition for evaluating classifier performance.

Table 1. Means and standard errors (SE) of reaction times, in seconds, for communication, dials and light tasks and mean error for the resource management tasks. The *F* values and probabilities for the comparison between the easy and difficult task levels are presented in the bottom row.

	Communication	Dials	Lights	Resource Management
Easy	2.52 (.42)	2.82 (.59)	1.69 (0.20)	495.25 (46.82)
Difficult	3.11 (.42)	3.78 (.59)	2.32(0.27)	747.65(95.69)
ANOVA	$F(1,7) = 15.24$ $p < 0.01$	$F(1,7) = 32.01$ $p < 0.01$	$F(1,7) = 71.66$ $p < 0.01$	$F(1,7) = 8.42$ $p < 0.02$

In order to assess the reliability of each of the classifiers in discriminating between easy and difficult task conditions using the physiological data, the mean proportion of correct classifications for the five days were examined. It was expected that the ability of the classifiers to generalize across days would increase as the number of days in the training set increased; presumably, the classifier learns those features that are reliable across the days in the training set. Figure 2 shows the accuracies obtained with each of the three classifiers as a function of number of days in the training set. The within-day accuracies are for the withheld test data from the same days as the training set (25% of the day's data set), while the between-day accuracies are for the days that were not part of the training set. Between-day accuracies have been averaged across all subjects and combinations of which days were in the training and test sets. All possible combinations of training and test days were permuted and tested, in order to reduce the impact of any one day being an outlier. For example, in the 1 Day condition, accuracies were evaluated using each of the five days separately to train a classifier, with the remaining four days combined to form the test set.

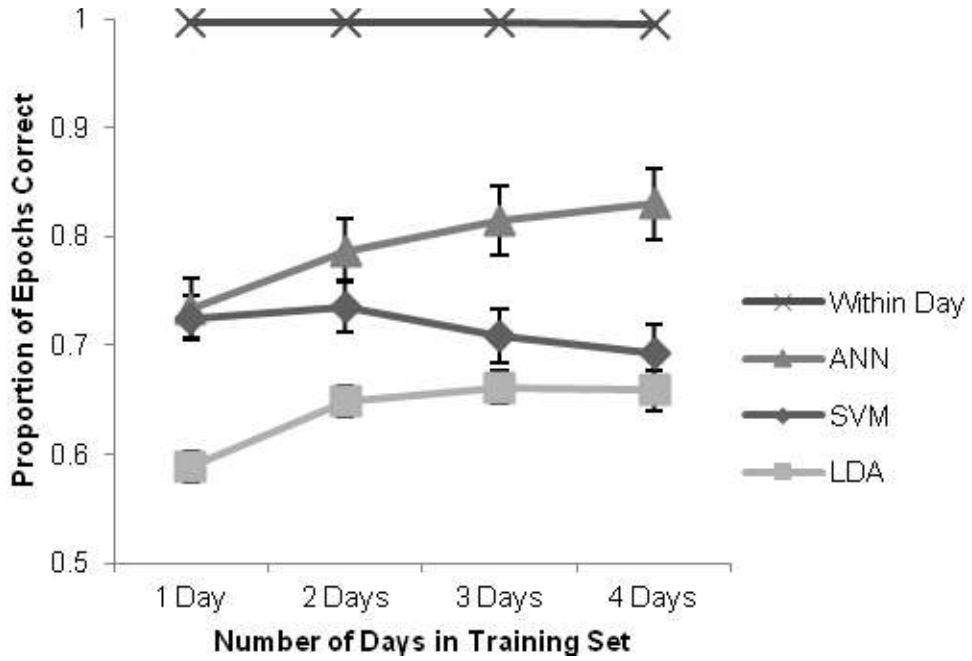

Figure 2. Classification accuracy as a function of method and number of days in the training set. Error bars in all figures are standard errors of mean proportion of epochs correct across subjects. The within-day average is collapsed across methods, as all three were at ceiling when trained and tested on the same day. Note that the test sets for the within-day proportion of epochs correct were randomly sampled (the withheld 25% test set) from all days that were combined to form the training set.

All three classifiers were at ceiling within-day, and have been collapsed for that condition. All three classifiers are also well above chance (.5) performance in all cases. However, the decrease in accuracy from within-day to between-day is substantial, amounting to at best a drop from .99 to .83 in the case where four days of data were used for training the ANN.

The SVMs were constructed using the LS-SVM formulation as well as an incremental method with leave-one-out validation. These methods produced very similar accuracies, with less than 1% difference on average across subjects. For simplicity, the reported accuracies are just from the LS-SVM. The SVMs were also constructed using either an optimized GRBF input kernel or a linear input kernel. Both kernels produced within-day classification at or near ceiling, however the linear kernel produced dramatically better between-day classification; consequently all reported SVM accuracies are from the linear kernel. This is generally consistent with previous results that demonstrated good generalization for linear SVM applied to magnetic resonance imaging data (Klöppel et al., 2008).

A three (classifiers) by four (days in the training set) repeated-measures ANOVA was performed on the between-day accuracies. There was a significant main effect of classifier, $F(2,14)=33.2$, $p<.01$, as well as a significant effect of days, $F(3,21)=8.2$, $p<.01$. The two-way interaction was also significant, $F(6,42)=9.2$, $p<.01$. The ANN produced both the highest overall accuracy as well as the largest increase with more days in the training set. The LDA did not perform as well as the ANN or the SVM. For the ANN, the prediction that increasing the number of days would improve between-day

accuracies was correct. Somewhat surprisingly, neither the SVM nor LDA showed such a consistent trend. It is possible that different approaches to implementing these classifiers could produce different results.

These analyses were conducted with the complete data set, including both EEG and peripheral physiological measures (ECG, EOG). In order to determine the relative contribution of peripheral measures to both accuracy and stability across days, this last analysis was repeated with only EEG features. Across conditions and classifiers, this resulted in a mean decrease in accuracy of 2%, with a range of 1.4 to 3.1%. Including peripheral measures improved classification accuracy generally, without increasing or decreasing stability across days. Consequently, all subsequent analyses include both EEG and peripheral measures.

The observed decline in classification accuracy across days could be caused by either poor generalization of the classifiers (overfitting), or by changes in the underlying distributions of data associated with the easy and difficult task conditions. The distributions associated with these classes were consequently examined post hoc; four examples are presented in Figure 3. Individual feature distributions are highly overlapped; however in the first two examples, there is a mean shift between the easy and difficult task conditions that reverses from the first day's data to the second. In the second two examples, the feature distributions are relatively stable across days. Based on these examples, it is unlikely that the decline in accuracy is solely a failure of generalization; without additional data that enables an assessment of feature stability, it is difficult to envision an *a priori* means of coping with a reversal of the ordinal relationship between classes.

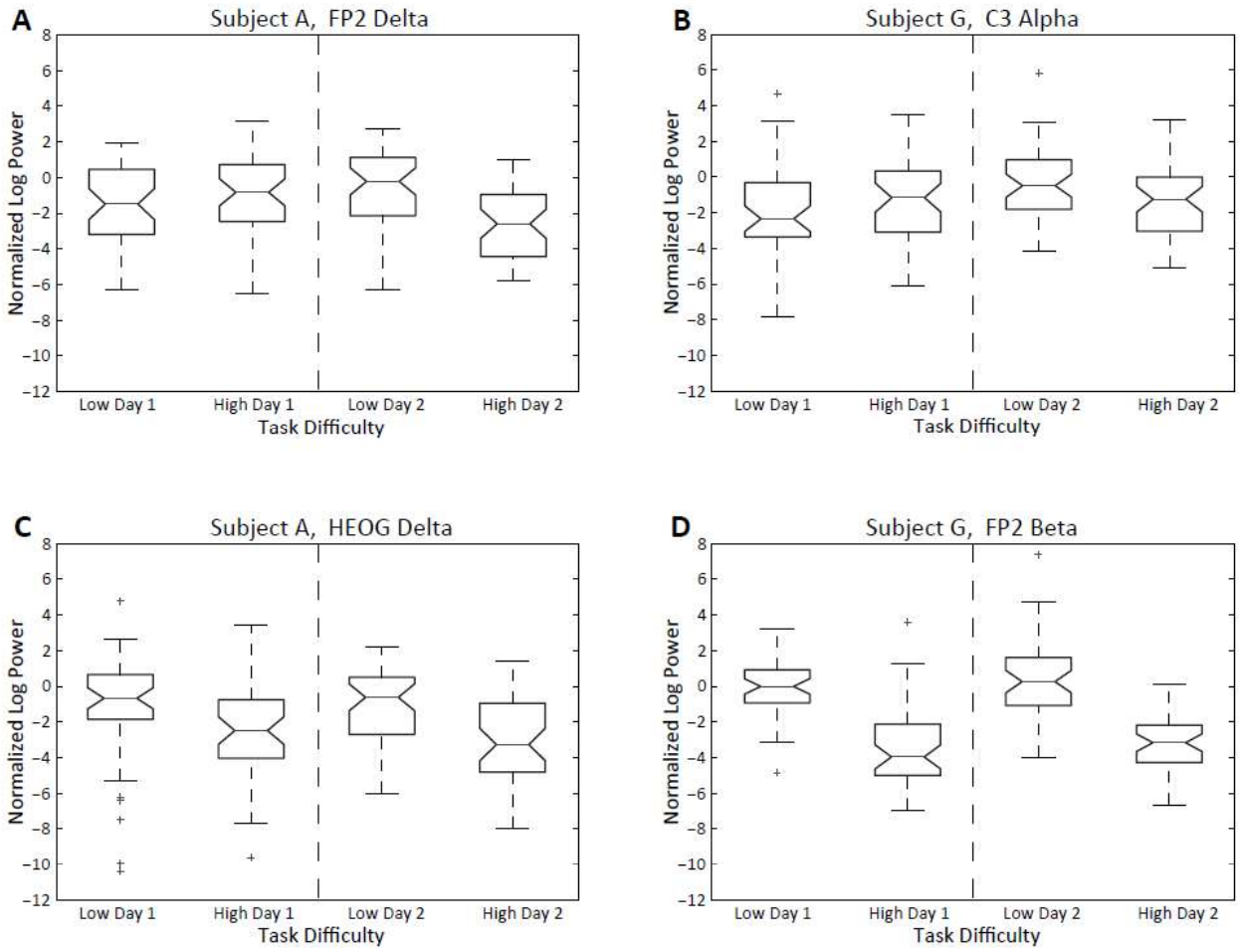

Figure 3. Sample feature distributions drawn from two subjects. Each panel plots the distributions for a selected feature and subject as a function of task difficulty and day. The vertical dashed line separates the first and second days of data collection. The boxes plot the mean and inner quartiles, with the whiskers extending to three standard deviations above and below the mean. Any outlier values beyond this range are plotted as crosses. Across subjects, any one feature exhibits relatively small differences between classes; however, 3A and 3B illustrate that these small differences can invert from one day to the next. 3C and 3D illustrate that features do exist that are more stable across days; presumably, improvements in classification accuracy associated with multiple days of training data involve weighting these features more heavily.

Knowing both that there is a significant decline in accuracy when classifying across days and that the ANN handled it best of our classifiers, additional ANN analyses were conducted to identify the time course of the decline in accuracy. Each five minute trial at a particular difficulty level was split into two halves, enabling comparisons (1) within halves (training and test sets seconds apart), (2) from the first to the second half of a trial (minutes later), (3) from one session to the next session of that day (hours later) and (4) from one data collection day to the next (days to weeks later). A first test was done to examine the effect of increasing the interval between days; there was no significant difference associated with increasing the interval from one day to one week to two weeks, $F(2,14)=1.08$, $p>.3$. Therefore, the results for the days in between have been collapsed across those intervals. The results of classifying across varying time periods are plotted in Figure 4. A repeated

measures ANOVA revealed a significant main effect due to the interval between training and test, $F(3,21)=17.9$, $p<.01$; the increase in time from seconds to hours resulted in a significant decrease in classification accuracy. It is possible that the accuracy in the seconds interval has been increased due to the overlapping window; non-overlapping forty-second windows would result in too little data to analyze. While slightly lower than hours on average, the accuracy between days was not substantially different. It appears that the primary decline in accuracy is on the scale of hours and does not change after up to a two-week interval; note that the two-week interval still resulted in classification accuracy across those days significantly above chance.

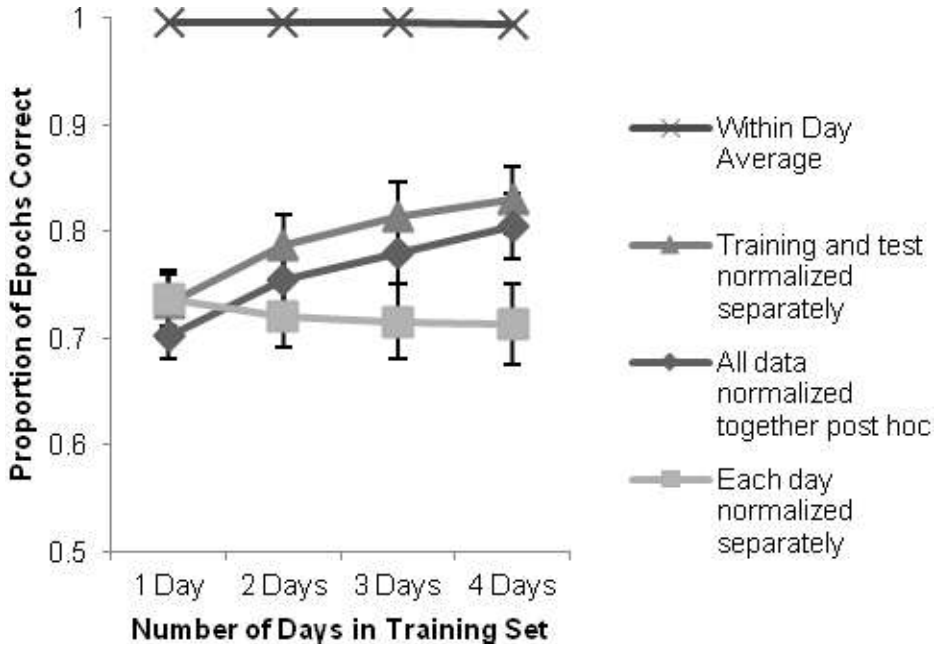

Figure 4. ANN classification accuracy as a function of time from training set. By subdividing each 5 minute trial into two halves, we obtained accuracy as a function of time between training and test sets: the reserved test data (25% witheld) from the training set is seconds apart, the second half of a trial is minutes apart, the first and last sessions are about 1 hour apart, and subsequent testing days are from 1 day to 2 weeks apart.

A practical solution to improving classification accuracy over time is to accept that unpredictable variability exists from day-to-day, and use small amounts of data from the beginning of a new day to retrain a classifier, thus including the day-to-day changes incrementally. This is essentially the same approach successfully used by Huang et al (2011) for their ERP data. This was accomplished iteratively with the ANN, using a training data set starting with one whole day's data, and then adding (in increments of one half-trial, or 2.5 minutes per data class) increasing amounts of data from a subsequent day. The test sets were then constructed as in Figure 4, drawing from the withheld training data, the next half-trial, the next session, or the next day. The prediction is that increasing amounts of training data from a new day should gradually improve classification accuracy as compared to testing a classifier trained only on the data from the previous day. The results of this analysis applied to the first and second days of data collection are plotted in Figure 5.

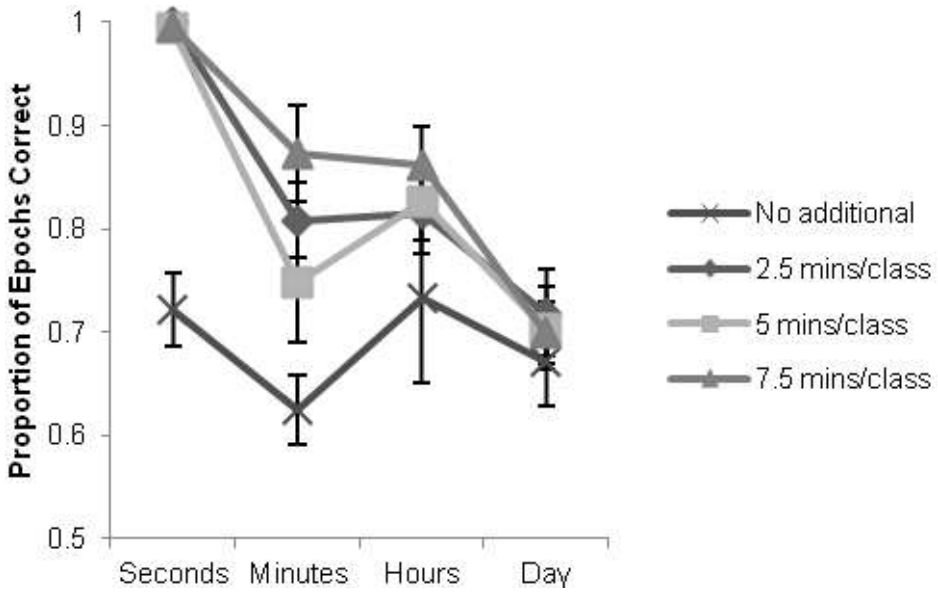

Figure 5. Classification accuracy as a function of quantity of data from a new day used to train the classifier. The training sets included all of Day 1 (No additional), and then increasing amounts of data from Day 2, taken in sequence (2.5 to 7.5 minutes per class). Test sets were from Day 2 (Seconds to Hours) or Day 3 (Day), with the time in between training and test sets given on the abscissa. "No additional" is a baseline calculated by testing a classifier trained only on Day 1 on the same test sets drawn from Day 2; please note that for this baseline, all test sets are separated from the training data by a day.

A four (from no additional training data through 7.5 minutes/class) by three (time between training and test, excluding the uninformative seconds category) repeated-measures ANOVA revealed a main effect of additional training data, $F(3,21)=8.8$, $p<.01$, a nonsignificant effect of time between training and test, $F(2,14)=2.5$, $p=.11$, and a significant interaction $F(6,42)=3.5$, $p<.01$. The interaction is likely significant due to the additional training data conditions being superior only at the minutes and hours levels. A planned comparison between the lowest (2.5 minutes per class) amount of additional training data and the baseline/no additional data condition resulted in significant improvement for the additional training data at the minutes level, $t(7)=3.0$, $p=.03$ (Bonferroni corrected), and marginally significant improvement at the hours level, $t(7)=2.3$, $p=.08$. The comparison for the days level was nonsignificant. Adding as little as 2.5 minutes of data per class from a new day resulted in improved performance over using only data from the previous testing data to train the ANN.

An additional consideration in analyzing data from multiple days is the source data for normalization parameters. If the object is real-time classification of new data, normalization parameters would have to be derived from the training set; on the other hand, if classification is being performed post hoc various options are available. We tested three logical alternatives: deriving normalization parameters separately for each training and test set, normalizing the entire data set together, and separately normalizing each day. This last method was intended to be representative of using calibration or small amounts of training data from each day to separately normalize. The results as a function of number of days in the training set are presented in Figure 6. The highest overall classification

accuracies are produced by separately normalizing training and test sets, as was done in the analyses presented above. Post hoc, mass normalization of the data set resulted in similar but consistently lower accuracies. Normalizing each day separately erased the benefit for including more than one day in the training set.

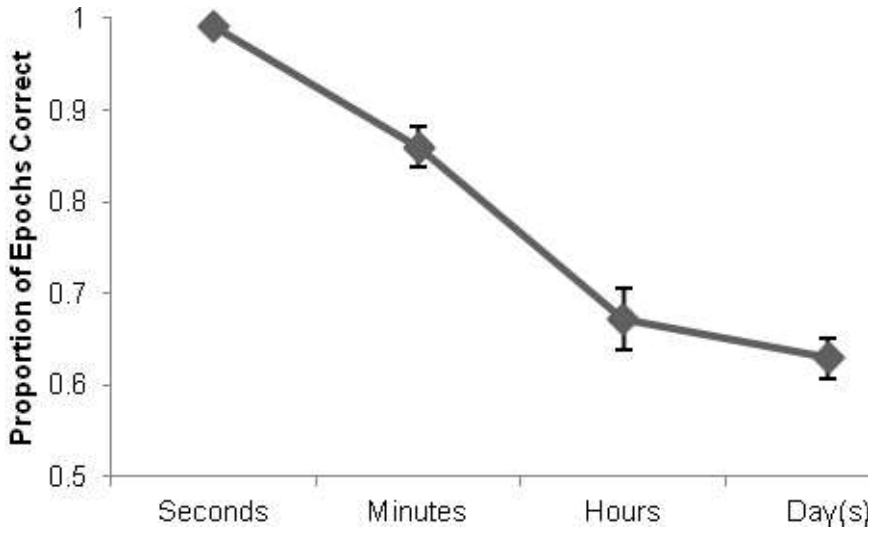

Figure 6. Classification accuracy as a function of number of days in the training set and normalization scheme. Separately normalizing each day to itself produced generally worse accuracy, while the other two schemes were not significantly different. The data presented in all the previous figures was generated by normalizing training and test separately.

5.0 DISCUSSION

The results of the present study replicate earlier reports by demonstrating that classifiers that are trained and tested on physiological data from the same day can very accurately determine which of two levels of task difficulty produced the data (Berka, et al., 2004; Freeman, Mikulka, Prinzel & Scerbo, 1999; Gevins, et al., 1998; Wilson & Fisher, 1991; Wilson & Russell, 2003a; 2003b). This, no doubt, contributed to the successful application of adaptive aiding using these procedures (Wilson & Russell, 2007). However, the results also show that the ability of the three classifiers to correctly classify easy from difficult task conditions deteriorates over time. In all cases, the accuracy levels remain above chance. However, improved classifier accuracies over multiple days would be beneficial in facilitating practical applications of adaptive aiding and other uses of operator functional state estimation. The present results suggest that the ANN classifier is superior to the SVM and LDA classifiers for this particular data set; it is likely that data sets with different structure would change their relative accuracies. Given that both the ANN and SVM with GRBF input kernel are nonlinear classifiers; it is somewhat surprising that the linear SVM performed better between days than GRBF while not outperforming ANN. One possible explanation for this difference is the handling of the validation set. Using the validation set as a check for overfitting in the ANN may have resulted in better generalization as compared to simply including that data in the training set as was done with the SVM and LDA classifiers. If that were the only cause, we would have expected SVM with leave-one-out crossvalidation to be closer to ANN performance. It is possible that further optimization of

any of these techniques would improve generalization across days, but the results obtained from the ANN are nevertheless encouraging that relatively high accuracies are achievable.

The within-day results for all three classifiers suggest that when the test data and training data are taken from the same larger data set, very high levels of test accuracy can be expected. All three classifiers produced nearly perfect discrimination of easy and difficult conditions by using the physiological data. However, when the test data were collected at a different time from training data, classifier accuracy declined. This was seen when the test data were from an entirely different day, and also when the test data were generated minutes and hours apart from the training data. Variations in the physiological data exist on a scale from seconds to days that reduce the ability of the classifiers to correctly identify the data. Variation in physiological data within a single day (circadian effects) has been extensively researched; however similar variation across multiple days has not been as well studied. The decline in classification accuracy appears to level off at the hours level and is maintained above chance from one day to up to two weeks later (Fig. 4). This should not be taken as indicative that the underlying causes of the observed variability are the same at the hours and days scale; Figure 5 demonstrates that hours and days can dissociate and may reflect different but roughly equal sources of variability. When data from our maximum of four days were added to the ANN training set, the accuracy levels across days reached the rather high level of approximately 83% correct classification (Fig. 2). This level of discrimination has been shown to be very beneficial when used to trigger adaptive aiding (Wilson & Russell, 2007). It is quite possible that these levels of classifier accuracy would be very useful in many situations requiring continuous estimates of OFS. This would be especially true of work environments where there is little or no performance data, such as in situations of high levels of automation where the operators primary responsibility is to monitor system functioning, detect outlier conditions and respond appropriately. Even being able to detect operator cognitive overload in hazardous situations with 83% accuracy would be very advantageous: as long as overall adaptive man-machine system performance under cognitive overload exceeds performance without OFS monitoring, such a system should be considered useful. Human operators gaining experience with such a system are also likely to adapt their own behavior to the reliability of the monitoring, helping to ameliorate the consequences of misclassifications. This could be particularly effective if the monitoring is implemented as a component of a decision aid (McGuirl & Sarter, 2006).

Rather than collecting multiple sessions and days of training data, another strategy to improve classifier accuracy is to adjust the classifier using relatively small amounts of physiological data from the current day. This would represent a compromise between the clearly effective but inefficient approach of training a new classifier each day, and doing no updating at all. The present results showed that adding training data from subsequent sessions improved the accuracy of the ANN classifier. By starting with one full day of training data and then adding from 2.5 min to 7.5 min of data per level of task difficulty from the test day, the accuracy of the classifier for the remainder of that test day was significantly improved. This result is consistent with Huang et al (2011), despite that work using ERPs associated with target detection, suggesting that the incremental addition of training data is a generally useful approach. The effect diminished when applied to an additional day after the test day from which additional training data was added (Fig. 5). This suggests that a practical solution to reliable OFS classification could be to conduct brief recalibration sessions each day following an initial longer training session. This sort of brief setup has the potential for incorporation in practical monitoring systems.

The underlying nature of the changes from day-to-day is not clear from the present results. The reversal of class ordering demonstrated in Figure 3 suggests that changes in the distributions of features associated with the easy and difficult task conditions is a likely contributor. Poggio et al. (2004) established that classifiers such as SVM may be provably stable and generalizable, however their analysis is predicated on the assumption that the data distributions (or generating functions) associated with each class are fixed. This may not hold for physiological signals associated with cognitive task difficulty; it appears that significant variations in distributions occur over time, including reversing ordinal relationships. This variation is not solely due to circadian effects within a day (e.g., Refinetti, 1999); otherwise accuracy should have been somewhat improved when testing with data from equivalent time periods on a subsequent day. The day-to-day differences in the physiological data negatively impacted all three classifier's accuracy results when tested on different day's data. However, very high levels of discrimination were found when the trained classifiers were tested on data from the same data collection days. This was true when the classifiers were trained on any of the single day's data. This suggests that there are characteristics in the physiological data that can be used by the classifiers to very accurately discriminate between easy and difficult task conditions on any given day.

The use of the different normalization schemes was an attempt to determine whether or not the variability from day-to-day could be ameliorated by making the data distributions more consistently normal by various methods (Fig. 5). This was unsuccessful and suggests that the solution to the differences between the physiological data from day-to-day is not likely to be just a matter of normalizing the data in different ways. As seen in the sample feature distributions, response characteristics of several of the 193 physiological features can dramatically change during complex task performance. If the order of classes within a feature is not preserved across a day, it is difficult to define an *a priori* means of assigning data to the proper class. Even if this is the case, the differences between classes are consistent within a given day or at least over several minutes if not hours of that day. Techniques that modify the data distributions such as unsupervised covariate-shift minimization (Satti, Guan, Prasad & Coyle, 2010) as well as unsupervised updating of classifiers are being developed for BCI in order to maintain discrimination accuracies; this has been recognized as a key challenge for deployment of BCI (Krusienski et al, 2011). Such techniques will likely prove directly applicable to mental state classification. Nevertheless, the current results showed that accuracy levels of 87% and 86% correct could be produced within minutes and hours (Fig. 4). With continued advancement in sensors, signal processing, and pattern classification, additional improvement can be expected that should further improve the practicality and real-world application of mental state classification.

6.0 CONCLUSIONS

This work unit demonstrated that the stability of workload monitoring via classification of neurophysiological data can be enhanced significantly, likely to a level sufficient for future applications. It could not address several of the key basic science questions, namely why this variability exists and what it may tell us about the ability of operators to perform tasks on different days. Future work, both basic and applied, should attempt to address these issues.

7.0 REFERENCES

Berka, C., Levendowski, D. J., Cvetinovic, M. M., Petrovic, M. M., Davis, G., Lumicao, M. N., Zivkovic, V. T., Popovic, M. V., & Olmstead, R. (2004). Real-Time Analysis of EEG Indexes of Alertness, Cognition, and Memory Acquired With a Wireless EEG Headset. *International Journal of Human-Computer Interaction*, 17, 151-170.

Birbaumer, N. (2006). Breaking the silence: Brain-computer-interfaces (BCI) for communication and motor control. *Psychophysiology*, 43, 517-532.

Blanco, J.A., Stead, M., Krieger, A., Viventi, J., Marsh, W.R., Lee, K.H., Worrell, G.A., & Litt, B. (2010). Unsupervised classification of high-frequency oscillations in human neocortical epilepsy and control patients. *Journal of Neurophysiology*, 104(5), 2900-12.

Burgress, A. & Gruzelier, J. (1993). Individual reliability of amplitude distribution in topographical mapping of EEG. *Electroencephalography and clinical Neurophysiology*, 86, 210-223.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York: Springer.

Cauwenberghs, G. & Poggio, T. (2001). Incremental and decremental support vector machine learning. In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.), *Advances in Neural Information Processing Systems 13* (pp. 409-415). Cambridge, MA: MIT Press.

Coburn, K.L., Lauterbach, E.C., Boutros, N.N., Black, K.J., Arciniegas, D.B., & Coffey, C.E. (2006). The value of quantitative electroencephalography in clinical psychiatry: a report by the Committee on Research of the American Neuropsychiatric Association. *The Journal of Neuropsychiatry and Clinical Neurosciences*, 18, 460-500.

Comstock, J. R. and Arnegard, R. J. (1992). The multi-attribute task battery for human operator workload and strategic behavior research. *NASA Technical Memorandum* No. 104174.

De Martino, F., Gentile, F., Esposito, F., Balsic, M., Di Sallea, F., Goebel, R. and Formisano, E. (2007). Classification of fMRI independent components using IC-fingerprints and support vector machine classifiers. *NeuroImage,* 34(1), 177-194.

Freeman, F. G., Mikulka, P. J., Prinzel, L. J. and Scerbo, M. W. (1999). Evaluation of an adaptive automation system using three EEG indices with a visual tracking task. *Biological Psychology*, 50, 61-76.

Garrett, D., Peterson, D. A., Anderson, C. W., and Thaut, M. H. (2003). Comparison of linear, nonlinear, and feature selection methods for EEG signal classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering,* 11(2), 141-144.

Gevins, A., Smith, M. E., Leong, H., McEvoy, L. K., Whitfield, S., Du, R. & Rush, G. (1998). Monitoring working memory load during computer-based tasks with EEG pattern recognition methods. *Human Factors*, 40(1), 79-91.

Haynes, J.-D., & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, 7, 523-534.

Huang, Y., Erdogmus, D., Pavel, M., Mathan, S., & Hild II, K.E. (2011). A framework for rapid visual image search using single-trial brain evoked responses, *Neurocomputing*, 74, 2041-2051.

Jasper, H. H. (1958). Report of the Committee on Methods of Clinical Examination. *Electroencephalography and clinical Neurophysiology,* 10, 370-375.

Kamitani, Y., and Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience,* 8(5), 679-685.

Klöppel, S., Stonnington, C. M., Chu, C., Draganski, B., Scahill, R.I., Rohrer, J.D., Fox, N.C., Jack Jr., C. R., Ashburner, J. & Frackowiak, R.S.J. (2008). Automatic classification of MR scans in Alzheimer's disease. *Brain,* 131, 681-689.

Kong, X. & Wilson, G.F. (1998). A new EOG-based eyeblink detection algorithm. *Behavior Research Methods*, 30 (4), 713-719.

Kriegeskorte, N., Simmons, W.K., Bellgowan, P.S., and Baker, C.I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nature Neuroscience*, 12, 535-540.

Krusienski, D.J., Grosse-Wentrup, M., Galan, F., Coyle, D., Miller, K.J., Forney, E. and Anderson, C. (2011). Critical issues in state-of-the-art brain-computer interface signal processing. *Journal of Neural Engineering*, 8 (2), 1-8.

Lal, T.N., Schroder, M., Hinterberger, T., Weston, J., Bogdan, M., Birbaumer, N., and Scholkopf, B. (2004). Support vector channel selection in BCI. *IEEE Transactions on Biomedical Engineering,* 51(6), 1003-1010.

Lippmann, R. P. (1987). An introduction to computing with neural nets. *IEEE ASSP Magazine*, 4-22.

McEvoy, L. K., Smith, M. E. & Gevins, A. (2000). Test-retest reliability of cognitive EEG. *Clinical Neurophysiology*, 111, 457-463.

McGonigle, D.J., Howseman, A.M., Athwal, B.S., Friston, K.J., Frackowiak, R.S.J., and Holmes A.P. (2000). Variability in fMRI: An Examination of Intersession Differences. *NeuroImage,* 11, 708-734.

McGuirl, J. M. & Sarter, N. B. (2006). Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Human Factors*, 48(4), 656-665.

Parasuraman, R., and Wilson, G. F. (2008). Putting the brain to work: neuroergonomics past, present, and future. *Human Factors,* 50(3), 468-474.

Pleydell-Pearce, C. W., Whitecross, B. T., & Dickson, B. T. (2003). Multivariate analysis of EEG: predicting cognition basis of frequency decomposition, inter-electrode correlation, coherence, cross phase and cross power. In R. H. Sprague (Ed.) *Proceedings of the 36th Annual Hawaii International Conference on Systems Science* (pp. 131-141) IEEE Computer Society. The Printing House USA.

Poggio, T., Rifkin, R., Mukherjee, S., & Niyogi, P. (2004). General conditions for predictivity in learning theory. *Nature,* 428, 419-422.

Pollock, V. E., Schneider, L. S. & Lyness, S. A. (1991). Reliability of topographic quantitative EEG amplitude in health late-middle-aged and elderly subjects. *Electroencephalography and Clinical Neurophysiology*, 79, 0-26.

Refinetti, R. (1999). *Circadian Physiology*. Boca Raton: CRC.

Salinsky, M. C., Oken, B. S. & Morehead, L. (1991). Test-retest reliability in EEG frequency analysis. *Electroencephalography and Clinical Neurophysiology*, 79, 383-392.

Satti, A., Guan, C., Prasad, G. & Coyle, D. (2010). A covariate shift minimisation method to alleviate non-stationarity effects for an adaptive brain–computer interface. *Proceedings of the 20th International. Conference on Pattern Recognition*, 105-8.

Shelley, J. & Backs, R. W. (2006). Categorizing EEG waveform length in simulated driving and working memory dual-tasks using feed-forward neural networks. *Foundations of Augmented Cognition.* Strategic Analysis, Inc. (pp. 155-161).

Smith, S.M., Beckmann, C.F., Ramnani, N., Woolrich, M.W., Bannister, P.R., Jenkinson, M., Matthews, P.M., and McGonigle, D.J. (2005). Variability in fMRI: a re-examination of intersession differences. *Human Brain Mapping,* 24, 248-257.

Suykens, J.A.K., Van Gestel, T., De Brabanter, J., De Moor, B. & Vandewalle, J. (2002). *Least Squares Support Vector Machines*. Singapore: World Scientific.

Widrow, B. and Lehr, M. A. (1990). 30 years of adaptive neural networks: Perceptron, madaline, and backpropagation. *Proceedings of the IEEE*, 78, 1415-1442.

Wilson, G. F., & Fisher, F. (1991). The use of cardiac and eye blink measures to determine flight segment in F4 crews. *Aviation, Space and Environmental Medicine,* 62, 959-961.

Wilson, G. F. & Russell, C. A. (2003a). Operator functional state classification using psychophysiological features in an air traffic control task. *Human Factors*, 45(3), 381-389.

Wilson, G. F. & Russell, C. A. (2003b). Real-Time Assessment of Mental Workload Using Psychophysiological measures and artificial neural networks. *Human Factors*, 45(4), 635-643.

Wilson, G. F. & Russell, C. A. (2007). Performance Enhancement in a UAV Task using Psychophysiological Determined Adaptive Aiding. *Human Factors*, 49, 1005-1019.

Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G. & Vaughan, T. M. (2002). Brain-computer interfaces for communication and control. *Clinical Neurophysiology*, 113, 767-791.

LIST OF ABBREVIATIONS AND ACRONYMS

ANN	artificial neural network
ANOVA	analysis of variance
BCI	brain computer interface
ECG	electrocardiography
EEG	electroencephalography
EOG	electrooculography
fMRI	functional magnetic resonance imaging
GRBF	Gaussian radial basis function
Hz	Hertz
IIR	infinite impulse response
kOhms	kiloohms
LDA	linear discriminant analysis
LS-SVM	least squares support vector machine
MATB	Multi-Attribute Task Battery
OFS	operator functional state
SE	standard error
SVM	support vector machine