A Systematic Approach to Developing Near Real-Time Performance Predictions Based on Physiological Measures

Amanda E. Kraft, Jon Russo, Michael Krein
Human Systems and Autonomy
Lockheed Martin Advanced Technology Lab
Cherry Hill, NJ, USA
{Amanda.e.kraft, Jon.c.Russo,
Michael.Krein}@lmco.com

Bartlett Russell, William Casebeer, Matthias Ziegler
Human Systems and Autonomy
Lockheed Martin Advanced Technology Lab
Arlington, VA, USA
{Bartlett.Russell, William.D.Casebeer,
Matthias.d.ziegler}@lmco.com

***Abstract*—Performance measurements using human sensing and assessment capabilities are limited by an inability to account for the multitude of variables that regulate performance state. Monitoring behavior alone is not adequate for prediction of future performance on a given task and no single physiological measurement can provide a complete assessment that influences performance. Here we investigate how to analyze a range of physiological measurements in near real-time using state of the art signal processing methods to predict performance.**

We developed multiple predictive computational models to assess when physiology markers that coincide with workload levels are reaching a point that performance decreases or increases may occur in the near future. Traditionally, models will vary significantly between studies (due to the diversity of tasks being tested, the number/type of sensors and differing analysis techniques), leading to specialized models that do not transfer between tasks and individuals. When a model is so specialized that it is only predictive to a specific task and not flexible to inter- or intra-individual differences without complete system retraining, it is impractical in applications outside of controlled experiments. To bring practical use of computational models in real world environments it is important to examine which types of physiological data that can both be reliably processed and analyzed in near real-time and that are highly predictive over time. It is also necessary to minimize the number of sensors in the real world so a sensor and signal sensitivity analysis needs to be performed.

We identified and collected physiological signals linked to workload including electroencephalogram (EEG), Heart Rate and Heart Rate Variability (HR/HRV) and Eye-Tracking while performing multiple tasks at varying difficulty levels. We tested a variety of preprocessing methods and computational models, including radial basis function kernel support vector machines and neural networks, to determine predictive power as well as computational time for each type of model. The models were tested using each signal independently as well as combinations of all the signals.

Keywords—Workload; Cognitive; EEG; HRV; predictive models

I. Introduction

Computational models have gained an increasing presence as techniques to understand human behavior and in linking behavior to physiological measures [1-3]. These models have successfully shown links between measures of workload with performance that were not previously apparent due to the large amount of data that current sensors are able to collect [4], [5]. While many studies implement such models, generalizing across studies is difficult due to diversity of task type and modified signal integration and processing techniques, as well as model implementation and parameter selection. Further, few studies evaluate models on multiple tasks, leading to highly specialized models that do not transfer between tasks and individuals. Consequentially, such models often require complete system retraining for usable predictions on novel tasks or individuals. To apply computational models outside of the lab with practical use in real world environments it is important to examine how physiological data can be reliably processed, what minimal number of sensors are necessary and how can the signals be analyzed in a manner that is beneficial for understanding workload levels and performance across both individuals and tasks of interest.

Studies have shown that cognitive workload levels can be measured using an increasing number of available sensing techniques. These measure of workload can then be used to track predict performance. Electroencephalography (EEG) has been one of the most common tools for measuring workload, identifying increased neural activity corresponding to workload levels [6-8]. Eye tracking and/or electrooculography (EOG) is also common, as evidence of pupil size and blink rate have been linked to workload levels [9], [10]. Electrocardiogram (ECG) offers another means to assess workload levels via heart rate variability [11]. By combining sensors some studies have been able to show workload levels consistent across multiple physiological sensors [5], [12] and to increase the classification accuracy of any one of these systems alone by accounting for a greater number of physiological systems that respond to changes in workload. In this study we use this combined sensing approach to measure performance in multiple tasks to determine how well general levels of workload are linked to task performance across individuals. Other tools, such as fNIRS, fMRI and biomarkers in particular [13], [14], have also shown to be important measures of workload we do not address these tools in this study, however we plan to in future work.

The ability of computational models to predict performance from physiological signals is an important tool that could have large number applications in cognitively demanding environments. Many studies have looked at linking workload with performance and have had success predicting performance [13], it is important to create a systematic and comprehensive study evaluating the possibilities and limitations of what a combination of physiological measures can predict. In this study, we start that process by looking at creating a single subject agnostic generalizable model to predict performance based on EEG, EOG, eye tracking and ECG and perform sensitivity analysis on each signal. We also don’t limit ourselves to a single model for a single task, but we compare accuracy of a model that is trained based on the performance results over two independent tasks and then tested on a separate hold-out population the model was not trained on. These results are compared with a model that is trained on a single trial from all individuals and tested on the same population over multiple additional trials as well as a model trained and tested on an equal, random distribution of all available data. We posit that subject’s physiological signals are unique enough that unless their performance is represented in the training algorithm there will be a decrease in model accuracy of performance prediction. However the exact tradeoff between drop-off in accuracy and individualization necessary for adequate performance is unknown. This study plays an important role in understanding difficulties that occur when trying to use physiological data as a measure of performance without individualized training.

Computational models can vary in complexity of programing, amount of data needed for training, time needed to run the model and number of parameters that need to be adjusted (i.e. layer size, learning rates, etc.) In this study we analyzed the results of two types of model, but focus primarily on one a deep belief network using neural networks. We chose a neural networking approach to model human performance, based upon the ability of neural networks to robustly classify nonlinear data. Additionally we did some initial testing comparing the neural network with a radial basis function kernel support vector machine (RBF SVM) model. In each model we did comprehensive cross-validation by randomly distributing the subject pool into training and testing sets and running the model 25 times to ensure accurate reporting the performance of each model. Simply running a model with a single training/testing set may cause skewed results as the training or testing data chosen may not be representative of the overall data. As performance variation between cross-validation runs is an indicator of dataset variability and an estimate for overall method reliability with respect to the data, we will present the models results we tested over the 25 model cross-validation runs and indicate that cross-validation should be standard procedure when testing models. We finally tested multiple combinations of sensors to see how eliminating them would change the predictive performance.

The result of our study shows that complex models do not always outperform shallow learning methods and that “more is better” does not hold up in collecting physiology features to predict performance on a task. Additionally, we show that some level of individualized tailoring is imperative for supporting the capability of accurately predicting performance changes in an individual never before seen. These findings provide an indication that a comprehensive study is necessary to understand the tradeoffs between generalized model performances versus the costs of adapting models to individuals.

II. METHODS

A. *Experimental Setup*

Participants. A mix of thirty-five right-handed undergraduate and graduate students from University of Maryland were trained and tested on two computer based video games to measure performance over a period of four non-consecutive days. Subjects were trained on the system during day 1 and tested days 2, 3 and 4. Four trials (each from different participants) were disregarded due to errors in recording.

Task Design. We designed two tasks to titrate workload: a simple Snake Game (see Fig. 1) and Prepar3D Flight simulator (Fig. 2), each of which contained multiple levels of difficulty. Subjects received one 45 minute training period to become familiar with the tasks and returned for 3 days following the training to perform the tasks. The order of the tasks and difficulty levels were randomized for each day. Each difficulty level lasted for 5 minutes in both tasks.

B. *Task Description*

Nokia Snake Game. A video game was developed using Presentation programming language to mimic the Nokia Snake game preloaded on Nokia cellular phones. The game,

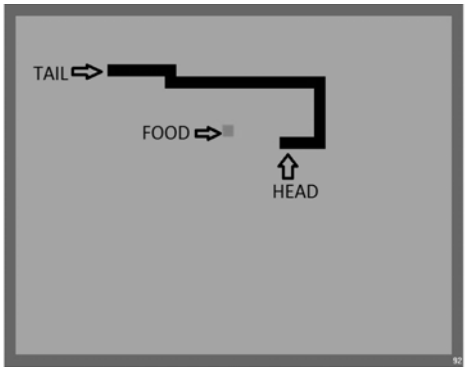

Fig. 1. Example of the “Snake Game”, the subjects control the head to eat the food while avoiding its own tail and the surrounding walls (grey boarder)

shown in Fig. 1, consists of a “snake” that moves at a constant pace, but the subject controls the snake’s direction with keyboard commands, including up, down, left or right. The goals of the game are to avoid hitting any walls or the snake itself (in which case the snake “dies” and the level restarts), and to collect as much “food” as possible. When subjects direct the snake to “eat” the food, the food adds a single additional square to the snake’s length There is no limit to the snake’s length, but as it grows longer, it becomes more difficult to navigate within the maze without hitting a wall or itself. The subjects completed two different levels of Snake: “easy” in which the snake moves at a slow speed (traveling across the screen only once every 100 ms) and a faster “hard” speed (moving once per 38 ms). The game provides an element of automatic titration to player skill; the length of the snake increases the difficulty of the subject’s ability to eat food while avoiding itself and the walls of the game. During testing we found that in the “hard” condition the keyboard input latency was delayed such that two fast-sequential keystrokes did not always register as the subject intended.

Prepar3D Flight Simulator. Lockheed Martin’s Prepar3D flight simulator was used as a second performance task. Subjects were given control of an aircraft and the task was broken down into 5-one minute subtasks. During minute one they were asked to maintain level flight at an altitude of 3000 ft. with a heading of 180 degrees at speed of 180 knots. In minute two, they were asked to maintain the same direction, but increase their altitude to 4000 ft while never increasing altitude at rate less than 1000 ft/min. They then maintained 4000 ft for another minute before decreasing again to 3000 and maintained that altitude for the final minute. To create multiple difficulty levels the environment in which the plane was flying was changed. In the “easy” condition there was no wind and no turbulence, however in the harder condition winds of over 30 knots and severe turbulence affected the aircraft’s position, causing high levels of difficulty maintaining the desired heading and altitude.

C. *Physiological Sensors*

During testing days subjects’ physiological signals were measured using BrainVision EEG system with electrodes arranged according to the standard 10-20 system [15]. Additional BrainVision electrodes placed on the collarbone recorded electrocardiogram (ECG) for HRV analysis. An SMI eye tracker recorded eye movements and pupilometry.

D. *Data Processing and Computational Models*

Model Development and Performance Estimation. We wanted to understand the impact of traditional signal processing methods on our ability to develop transferable human performance models. We baseline corrected the EEG data using the average Welch’s Power Spectral Density (PSD) computed from each individual’s resting state “eyes-open” session and removed the data for the first and last 10 seconds for each trial. We computed the PSD for each task, using Welch’s method over 1 second interval of data. For each channel and frequency bin, the corresponding baseline PSD average was subtracted from the task PSD. Each resulting 1-second interval was used as a unit of data for training or testing during model development and validation.

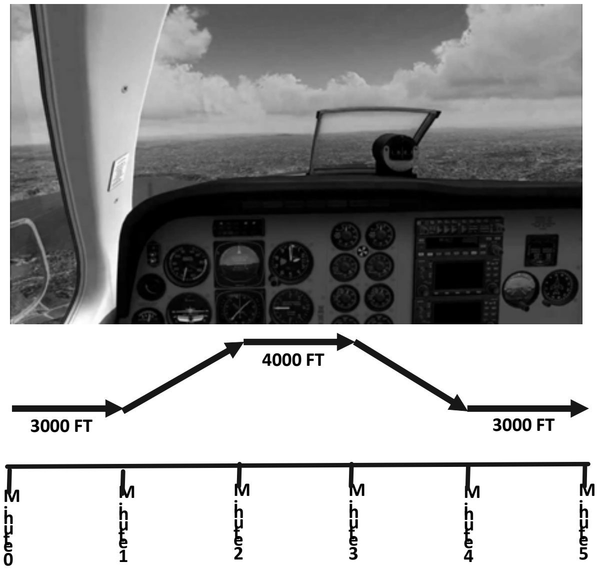

Fig. 2. Example seen from Prepar3D cockpit (left) task instructions (right)

The data was tested in two models a RBF SVM model (shallow model) and a neural network model (deep belief network). The deep belief network structures were Gaussian-Bernoulli Restricted Boltzmann Machine (GB-RBM) classifiers based upon Masayuki Tanaka’s code [16]. We estimated appropriate learning rates, learning step sizes, hidden layer sizes, and drop rates based on successive modeling performance during automated model tuning. We tuned models via sequential grid tuning approach; for a particular layer size and drop rate, a grid of step rates (0.1, 0.2, 0.3….) was evaluated. A step rate is then selected based on highest mean AUC and is used for subsequent modeling. The process is repeated for determining drop rates (0, 0.25, 0.5, 0.75…). For all models developed, the layer size chosen for use in determining the optimal learning rate was 100; the drop rate used for selection of optimal learning rate and hidden layer size was 0.5. All human data were standardized via z-score scaling prior to modeling.

Data Analysis. For each of the preprocessing strategies outlined above, Receiver Operating Characteristic (ROC) curves were analyzed and areas under the ROC curve (AUCs) are reported. An AUC of 0 refers to no prediction ability of the model, 0.5 denotes chance prediction, and 1.0 denotes perfect prediction of the model. Here we report the average and standard deviation of AUCs based upon 25 rounds of cross validation. The predictions are based on a binary decision for the model. The model predicts if the current test data that is provided to the model is above or below the median score for the task.

Cross Validation. In one of our approaches to cross-validation, we randomly assign 75% of the subjects (all trials) as a training set, and withhold the remaining 25% of subjects as a validation set. This is key to the approach, and reflects the transferability of human performance indicators without prior knowledge. It is important to note that since a developed model here has not been trained on any prior data from the validation set individuals, one would expect modest predictive ability. We compared this to other modeling strategies where data were withheld such that some data from each individual was included in both the training and testing set. We tested this by withholding a single trial from each individual randomly for the validation set (and the rest of the trials were part of the training set). In addition, we performed training modeling experiments where a subset of all subjects trials were included in the validation subject's, the balance of data (75%) remained in the training set.

Sensitivity Analysis. After the model results were tested we re-ran the best performing models with either a single set of signals from a particular sensor (i.e. eye tracking), a single EEG electrode placement or a single EEG signal band to determine the role of each signal in the broader predictive models

III. RESULTS

A. *Neural Network Models*

The neural networks showed significant differences in performance based on the type of model used (parameters chosen), type of training performed within the model and the specific task, data not shown. The performance also differed based on the type of task being performed, Fig. 3. The neural network models performed with higher accuracy 0.80 +/- 0.01 performance prediction accuracy when predicting the Prepar3d task as opposed to 0.76 +/- 0.01 during the Snake task.

B. *RBF SVM Model*

The RBF SVM model also showed difference in predictions between the types of tasks The performance also differed based on the type of task being performed, Fig. 3. The neural network models performed with higher accuracy 0.81 +/- 0.01 performance prediction accuracy when predicting the Prepar3d task as opposed to 0.79 +/- 0.01 during the Snake task.

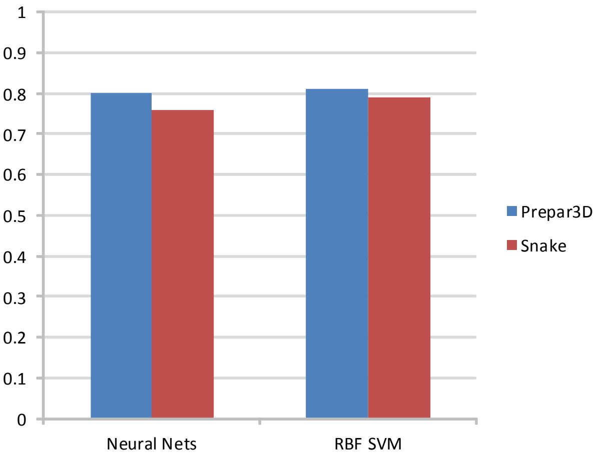

Fig. 3 Model Performance by Task

C. *Sensitivity Analysis*

After testing with a full set of data we systematically tested how each signal independently performed in predicting a subjects performance level. The results, shown Fig. 4, show that some signals have higher prediction performance measures (Beta Band of the EEG signal 72%/68%) than others which were relatively non-predictive (EOG blink 58%/51%). The sensitivity analysis did show that no single physiological signal could individually predict the performance that a combined system measures.

IV. DISCUSSION

As science pushes to obtain as much physiological data as possible in order to understand how human performance is holistically defined, they will rely on more complex computational models. This "more is better" approach may be beneficial for investigative purposes, but as we collect data we must systematically identify the best modeling approaches and eliminate those signals and sensors that do not contribute to predictivity. The results of the experiments in this study show that how the data modeled can create large variations in overall predictive power. In order to choose the correct computational model and the correct signals, one must take multiple factors into consideration, which we will discuss here.

A. *Physiological Sensors*

In this experiment we chose multiple physiological sensors that would record measurements that have been linked to cognitive workload levels. From modeling on individual feature types, we showed that EEG features contributed significantly more to predictivity than ECG, EOG, and eye tracking features combined. Although the number of inputs into the model were not equal (e.g. n=26 for Beta Mid and n=19 for non-EEG feature inputs as shown in Fig. 4), the models were independently tuned based on input. Hence it is more likely that the signals themselves do not convey appropriate information from which the model can learn, rather than a limitation due to model parameter specification. We are currently working on integrating different features derived from ECG signal that might be more sensitive to autonomic balance and drive, as well as adapting alternative methods for blink analysis in EOG signal. As adding more features increases the computational burden, we have also looked at feature reduction. Preliminary results from models in which EEG data is separated into broad bands (i.e. Alpha, rather than Alpha High and Alpha Low) suggest that the additional information gained is negligent in light of the computational burden in processing and modeling on double

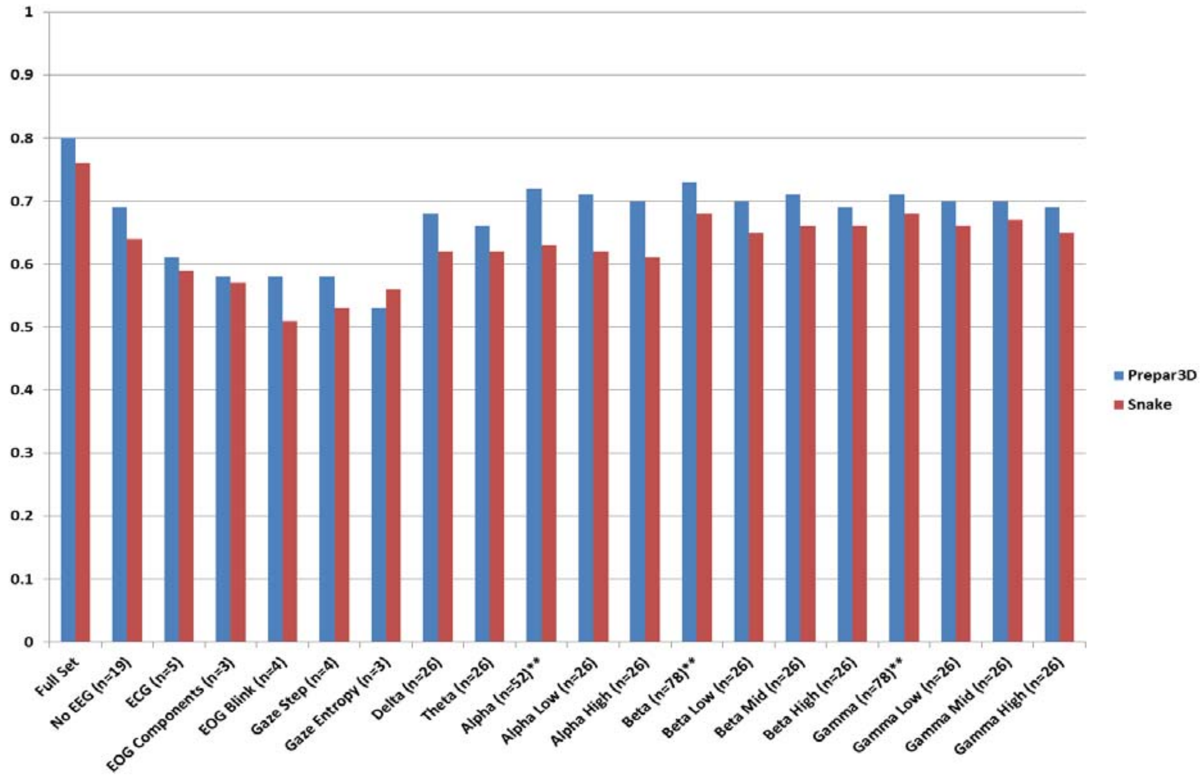

Fig. 4 Sensitivity Analysis

computational double the features.

Another aspect of signal processing that future studies should evaluate is the temporal alignment of physiological and behavioral data. When using modeling approaches that only associate inputs and output of a single observation, it is critical to ensure the data is synchronized in time. Given synchronized data, it is important to establish the appropriate integration window size per signal type and the delay of signal change manifesting in behavior. For example, cortical signals fluctuate on the order of milliseconds, while HRV is on the order of seconds. Further, precise timing may vary depending on what modalities and/or level of processing is necessary to complete a task. Alternative models capable of learning patterns over a series of contiguous inputs may offer a way to compensate for this issue. However, to avoid a “black box” approach, such a model would need to be decomposable enough to allow insight on its derived temporal parameters and corresponding contribution value, such as rate of change per input type.

B. *Task Design*

Our original predictions were that workload would be low and performance high during the easy conditions, and as the task became more difficult the workload would increase and performance would fall below the median. However a one-to-one relationship does not exist between workload and performance, as has been demonstrated in studies comparing workload and performance in individuals with varying levels of expertise [13]. To account for this, subjects were screened for having little to no experience with either type of task and received an equal amount of time to train on each task. The difficulty levels of the tasks were also designed to require drastically different levels of workload. For example, during the easy flight simulator condition, to maintain constant altitude, heading and speed the subject simply needed to hold the flight stick steady and pull back with minimal force to change altitude. For the hard level, independent of expertise, large effort and attention was needed as the winds and turbulence changed the planes course affecting all performance levels agnostic to how well the subject controlled the plane. Similarly the Snake game speed on the easy level was slow enough that avoiding the walls was a simple task and subjects were able to control the snake and comfortably increase its length without much pressure. When switched to hard, the subjects not only had to increase concentration based on the speed of the snake, but as an unattended consequence of our task not recording every keystroke, subjects had to change strategies in real time to compensate for keystrokes not responding, theoretically increasing workload levels. The performance graphs of the easy and hard conditions of each task showed distinct distributions in performance. While the trend comparing model performances was consistent across tasks, there were significantly greater differences in the flight simulator task over the snake task. This difference may be attributed to multiple discontinuities between the two tasks.

While the flight simulator had distinct quantitative goals (altitude, heading, speed, feet/minute) that every subject aimed for, the snake task allowed for a subjective discretion if the priority was to stay “alive” or to eat as much food as possible to obtain the same performance measure. Second, the altered controls in the hard snake task may have caused subjects to adopt a different strategy in how they approached the food. For instance, in the easy task subjects could make quick turns to avoid the wall or obtain food, but in the hard task they might opt for larger turns to compensate for the delay in response to keystrokes, thus lowering their score or potentially frustrating some subjects, causing workload changes independent of the task difficulty levels.

C. *Training Method*

When we trained both the deep belief networks and the RBF SVM on a 75% subset of the subjects and tested using the remaining 25% of the subjects the models performed only at a ~50% prediction accuracy (validation). Under our 25 model runs, altering which subjects were part of the training and which part of the testing the best results were 65% for the RBF SVM and 59% for the Deep Belief Networks. These results showed that while some distribution of the subjects caused the model training to be more representative of the larger population, the models were still poor at predicting performance based off of workload measures. The fact that multiple models showed this performance led us to believe that even with similar performance, there was not a standard workload measure that worked consistently across individuals for either task.

The model was tested for its ability to accurately predict performance from individualized workload measures. To do this, we trained the deep belief networks in a number of other ways to determine if we could generate a better performing model. The trained the models were tested using a subset of each subjects’ data the models performance significantly improved for the Deep Belief Network, but we saw no improvement in the RBF SVMs. First we tested by assigning a single session of each subject to train and tested on the remaining sessions (both models), second we randomly chose data points from each subject across all of their data (deep belief only) allowing each subject and testing session to be represented in both the training and testing model runs. Both model runs improved the performance of the DBN, the later showed the greatest improvement accuracy predicting up to ~80% of the tests in the flight simulator task. The worst cross-validation model performance with this modeling technique was equivalent to or only slightly better than the best general model where there was no training/testing overlap. The model’s improved ability to accurately predict performance form physiological workload measures when all subjects are represented in both the training and testing sets illustrates the extreme differences within physiological measures across individuals corresponds to the same behavioral outcomes. Only when a model is personalized for the intended user and possibly to a specific task, will it be reliable and useful as a predictive model.

We posit that for models to perform all participant sessions need to be represented in the training session. This is due to learning that may occur within subject across sessions. It is not only necessary to account for individualized differences when developing a computational model to predict performance or categorize workload, but a model most also account for learning that occurs over time. It has been shown that even experts in a given field have been shown to change performance and continue learning, albeit at a slower rate, thus models must account for this even in cases when naïve subjects are not being used. A possible method around this is to find subjects who may have had similar performance learning curves and similar physiological workload measures. Using a combination of the current users model and “future” models from the similar minded individual.

This may be extremely complicated though and will require future work. This work will examine the ability to group subjects based off of current and prior performance and create a set of template models to which a subject can be quickly matched. The template may then be tailored to the individual as the subject improves at the task shortening the overall process. By having not one, but a set of models trained on only subjects that show the same performance trends we posit a high accuracy prediction without a one-to-one relationship between number of models and subjects. This set of models will be the only possibility to create real-time modeling that will be necessary if adjustments to the subjects’ performance or tasking are desired in timely fashion.

ACKNOWLEDGMENT

We would like to acknowledge Bradley Hatfield, Li-Chun Lo, Hyuk Oh and Kyle Jaquess at the University of Maryland for their assistance in collecting data and technical support. Also we would like to acknowledge Ian Warfield of Lockheed Martin for his work in task development

REFERENCES

[1] D.E. Kieras, and Da. Meyer, “Computational Modeling of Human Multiple-Task Performance,” No. TR-05/ONR-EPIC-16. Michigan Univ, Ann Arbor Dept of Electircla Eng. and Comp. Sci. (2005)

[2] J.Hugo, and D.I. Gertman, “The use of Computational Human Performance Modeling as Task Analysis Tool,” In Proceedings of the Eighth American Nuclear Society International Topical Meeting on Nuclear Plant Instrumentation, Control, and Human-Machine Interface Technologies, NPIC&HMIT 2012 (pp. 22-26)

[3] J. Meng, X. Wu., V. Morozov, V. Vishwanath, K. Kumaran and V. Taylor, “SKOPE: A framework for modeling and exploring workload behavior,” In Proceedings of the 11th ACM Conference on Computing Frontiers , 2014

[4] Y. Ke, H. Qi, F. He.,S. Liu., X. Zhao,P. Zhou and D. Ming, “An EEG-based mental workload estimator trained on working memory task can work well under simulated multi-attribute task,” Frontiers in human neuroscience, 8, 2014

[5] Y. Liu, H. Ayaz.,B. Onara and P.A. Shewokis, “Neural Adaptation to a Working Memory Task: A Concurrent EEG-fNIRS Study,” In Foundations of Augmented Cognition, Springer International Publishing, 2015, pp. 268-280

[6] A.T. Kamzanova, A.M. Kustubayeva and G. Matthews, “Use of EEG workload indices for diagnostic monitoring of vigilance decrement.

Human Factors," The Journal of the Human Factors and Ergonomics Society, 56(6), 2014, pp. 1136-1149

[7] C. B Walter, "EEG workload prediction in a closed-loop learning environment" [Doctoral dissertation, Universität Tübingen], 2015

[8] A. M. Brouwer, M.A. Hogervorst, J.B. Van Erp, T. Heffelaar, P.H. Zimmerman and R. Oostenveld, "Estimating workload using EEG spectral power and ERPs in the n-back task," Journal of neural engineering, 9, 2012

[9] I. P Bodala, S. Kukreja, J. Li, N.V. Thakor and H. Al-Nashash, "Eye tracking and EEG synchronization to analyze microsaccades during a workload task," In Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE, 2015, pp. 7994-7997

[10] B.Zheng, X. Jiang, G Tien, A Meneghetti, O.N.M. Panton and M.S. Atkins, "Workload assessment of surgeons: correlation between NASA TLX and blinks," Surgical endoscopy, 26(10), 2012, pp. 2746-2750

[11] Y. Ke, H. Qi, F. He, S. Liu, X. Zhao, P. Zhou and D. Ming, "An EEG-based mental workload estimator trained on working memory task can work well under simulated multi-attribute task," Frontiers in human neuroscience, 8, 2014

[12] J. Choe, B.A. Coffman, D.T. Bergstedt, M.D. Ziegler and M.E. Phillips, Transcranial direct current stimulation modulates neuronal activity and learning in pilot training. Frontiers in Human Neuroscience, 10, 2016.

[13] H. Ayaz, P.A. Shewokis, S. Bunce, K. Izzetoglu, B. Willems, and B. Onaral, "Optical brain monitoring for operator training and mental workload assessment," Neuroimage, 59(1), 2012, pp. 36-47.

[14] M.A. Just., P.A. Carpenter and A. Miyake, "Neuroindices of cognitive workload: Neuroimaging, pupillometric and event-related potential studies of brain work,". Theoretical Issues in Ergonomics Science, 4(1-2), 2003, p. 56-88.

[15] H.H. Jasper, "Report of the committee on methods of clinical examination in electroencephalography: 1957". Electroencephalography and Clinical Neurophysiology 10 (2), 1958, pp. 370–375

[16] M. Tanaka and M. Okutomi, "A novel inference of a restricted boltzmann machine," In 2014 22nd International Conference on Pattern Recognition (ICPR) IEEE, 2014, pp. 1526-1531