Training ML Classification Models of Warfighter State with fNIRS

Olivia Fox Cotton, Justin Morgan, Lisa Lucia, Will Dupree, Jordan Coker, Matthew Ewer
Aptima, Inc.
Dayton, Ohio
{ofox, jmorgan, llucia, wdupree, jcoker, mewer}@aptima.com

LCDR Joseph Geeseman
Naval Air Systems Command
Patuxent River, Maryland
Joseph.w.geeseman.mil@us.navy.mil

ABSTRACT

Information regarding the state of Warfighters (e.g., workload, fatigue, distraction) can greatly impact outcomes in training contexts and the likelihood of mission success in operational environments. Harnessing neural data collected via functional near-infrared spectroscopy (fNIRS) can be a game-changing tool for assessing individual and team states. However, most neural measurement devices are plagued by barriers of cost, size, portability, and ease of use by non-experts. The purpose of this research was to assess whether low-cost, LED-based fNIRS devices can appropriately classify operator workload states in varied task difficulty conditions. A lightweight, easy-to-adopt fNIRS system was developed that translates raw neural data into estimates of functional states by providing access to cerebral oxygenation data from the prefrontal cortex (PFC). These data can indicate a variety of executive functioning activities relevant to operators, including attentional focus, inhibitory control, and stress control, among others. Behavioral and physiological task data, and self-report data (NASA Task Load Index) were collected, with participants engaging in an updated version of the synthetic work environment (SynWin) called Aviator SynWin. Like its predecessor, Aviator SynWin is a computer-based task environment that requires participants to simultaneously perform four unrelated tasks. Additionally, participants completed the n-back task, the Continuous Performance Test (CPT), and the Flanker Task to provide calibration data to train models of operator state such as workload. Individualized models of workload classification were trained for each participant, and while classification accuracy was high within tasks, the models failed to classify workload at high levels of accuracy when attempting to generalize across tasks. Performance scores were negatively correlated with self-report workload across tasks, but significant correlations with performance were not observed when examining trial averaged HbO data.

ABOUT THE AUTHORS

Olivia Fox Cotton, M.S. is a Scientist at Aptima, Inc. with a background in the fields of human factors psychology, cognitive neuroscience, and cognitive systems engineering. She holds an MS in human factors and industrial organizational psychology from Wright State University, and a BA in psychology from Clemson University. She is currently pursuing her doctorate at Wright State University where she performs research in cognitive neuroscience.

Justin Morgan, M.S. is an Associate Researcher Engineer at Aptima, Inc. with 3+ years of experience in human research, data science, and programming in R, Python, Flutter, and C. He has supported projects that involve physiological sensing, behavioral data, and algorithm development. He holds an M.S. in HF/IO Psychology from Wright State University and a B.S. from Eastern Kentucky University, with a major in Psychology, concentration in the Brain and Cognitive Sciences, and minors in Computer Science and Mathematical Sciences.

Lisa Lucia, Ph.D., Senior Scientist, Aptima, Inc. works within the Intelligent Performance Analytics Division. There, she leads the Sensor-based Assessment Technologies capability, a portfolio of approaches and solutions that leverages human cognitive, physiological, and physical activity data for state assessment and predictive analytics that guide decision making and inform interventions in both training and real-world contexts. Dr. Lucia holds a PhD and an MS in cognitive neuroscience and experimental psychology from Tufts University and a BS in biological psychology from Bates College. She maintains Project Management Professional (PMP) and ScrumMaster® certifications and has been awarded Value Creation Practitioner and Mentor badges from Northeastern University's D'Amore-McKim School of Business.

Will Dupree, Ph.D. serves as a Sr. Research Scientist and Data Science Team Lead for Aptima, Inc.'s Intelligent Performance Analytics Division. He's led and contributed to multiple research efforts for Department of Defense agencies such as the National Geospatial-Intelligence Agency, Air Force, and Army, leveraging his research skills in the domains of machine learning (ML) and artificial intelligence (AI). His research interests include deep learning for time series analysis, network graph modeling, and causal inference.

Jordan Coker, M.S. is a Research Engineer with a master's degree in mechanical engineering, specializing in biomechanics. He focusses on data processing of physiological and biomedical signals, developing innovative techniques that leverage machine learning to predict and understand human intent and capabilities. He has also contributed to the creation of physical products, including advanced insoles and exoskeletons, designed to enhance human performance in various tasks.

Matthew Ewer, M.S. is a Senior Software Engineer at Aptima, Inc. and has worked on project teams focused on the measurement of physiological state, operator health and safety, and operator performance for both teams and individuals. He has spearheaded the ongoing development of the SUPERGAK, a generalized software development kit (SDK) for future such projects. Mr. Ewer holds a master's degree in computational mathematics from Stanford University's Institute for Computational and Mathematical Engineering and a BA from Cornell College, with a major in mathematics and computer science.

LCDR Joe Geeseman, Ph.D. is a US Navy Aerospace Experimental Psychologist (#148) with a PhD in Brain & Cognitive Sciences from Southern Illinois University - Carbondale. Currently serving as the Military Deputy in the Naval Aviation Training Systems and Ranges Program Office (PMA 205), he manages a complex program integrating live, virtual, and constructive (LVC) aircraft systems for global operational training. His diverse experience includes multiple program and project management roles across NAVAIR, NAWCAD, CNATRA, and CDAO; a visiting scholar at Stanford University where he worked on autonomous driving, fNIRS, and stress responses in humans; and he is consistently overseeing multimillion-dollar research budgets to modernize Naval Aviation with cutting-edge technologies.

Training ML Classification Models of Warfighter State with fNIRS

Olivia Fox Cotton, Justin Morgan, Lisa Lucia, Will Dupree, Jordan Coker, Matthew Ewer
Aptima, Inc.
Dayton, Ohio
{ofox, jmorgan, llucia, wdupree, jcoker, mewer}@aptima.com

LCDR Joseph Geeseman
Naval Air Systems Command
Patuxent River, Maryland
Joseph.w.geeseman.mil@us.navy.mil

BACKGROUND

The ability to optimize the speed and quality of training for military personnel supports the US military's goal of maintaining land, sea, and air superiority. Accessing information regarding the state of Warfighters (e.g., workload, fatigue, distraction) can greatly impact outcomes in training contexts and the overall likelihood of mission success in operational environments. For example, Cognitive Load Theory (CLT) suggests that cognitive state is a key component of learning and that ideal learning conditions require maintenance of an optimal level of cognitive load without inducing over- or underload (Sweller, 1988). Similarly, it has long been shown that achieving an ideal level of attentional arousal has been linked to better performance outcomes (Yerkes & Dodson, 1908). While numerous subjective approaches to accessing relevant state information have been developed and utilized, such solutions require interruptions to workflows to gather data and are subject to potential response errors (e.g., bias, careless responding). In contrast, physiological data offers an objective alternative for passively accessing state information.

The phenomenon of physiological responses induced by changes to operator states (e.g., cognitive workload) is also well established in the literature (e.g., Ranchet et al., 2017; Hancock & Matthews, 2019). Harnessing neural data collected via functional near-infrared spectroscopy (fNIRS) can be a game-changing tool for assessing individual and team states in military contexts. The emerging technology of fNIRS is noninvasive, easier to field than electroencephalography (EEG), and has experienced significant technological advancement over the last two decades (Pinti et al., 2020). Modern fNIRS solutions are smaller, more portable, and wireless, all while retaining the ability to access cerebral oxygenation data from the prefrontal cortex (PFC). Operating on the assumption that task demands elevate metabolic activity and lead to a corresponding increase in frontal blood oxygenation, these data can indicate a variety of executive functioning activities relevant to operators, including attentional focus, inhibitory control, and stress control, among others (Fuster, 2015).

One potential drawback of physiological data is that quality sensors acquire data at high sampling rates, resulting in a large volume of data that can be difficult for researchers to efficiently manage and utilize in a timely manner. To truly achieve online state monitoring, advanced data modeling approaches must be employed to allow for automatic and real-time processing, analysis, and utilization of physiological information. Machine learning classification models constitute an ideal and optimized use of the high volume of data associated with wearable sensors.

One specific type of supervised machine learning that may be promising for application to physiological data is the use of long-short term memory (LSTM) models. LSTMs are a type of recurrent neural network (RNN) architecture that is particularly well-suited for handling timeseries data. Unlike traditional RNNs, LSTMs are designed to learn long-term dependencies in the data, therein overcoming the vanishing gradient problem. LSTMs achieve this by introducing a specialized cell state and a set of gates that control the flow of information into and out of the cell state. As a result, they can selectively remember and forget relevant information from previous time steps, making them highly effective at capturing temporal patterns and dependencies in timeseries data. The LSTM's ability to maintain a long-term memory and its sensitivity to both short-term and long-term patterns make it a powerful tool for modeling timeseries data such as physiological data streams.

As with other modeling approaches, generating LSTMs requires training data, validation data, and test data. Training LSTMs using data obtained from a standard cognitive task that is known to influence workload offers the potential benefit of producing a generalizable model that could be applied to a range of operational tasks where

workload ground truth may not be accessible in advance. The n-back task (Kirchner, 1958) is a potential source of appropriate data for such modeling activities, as there are many examples in the literature demonstrating the successful use of the n-back to quantify cognitive workload (e.g., Brouwer et al., 2012, Berka et al., 2007). The n-back requires participants to view a sequence of stimuli such as letters or pictures, and for each stimulus presented in the sequence, a participant must judge whether it matches the one presented *n* stimuli ago. Varying the *n* allows researchers to capture the workload associated with the processing of memory in differing demand conditions. Since the PFC is critical for the processing of memory and the workload associated with such activities (Herff et al., 2014), and fNIRS devices are capable of reliably capturing cerebral oxygenation of the PFC, the n-back is an ideal calibration tool for generating workload models that could be deployed in military training settings.

The goal of the present study was to train individualized models of participant workload using fNIRS data obtained while participants perform the n-back task, and then test the resultant models on data obtained from a task called Aviator SynWin that approximates the demands of a simulator-based aviation training environment. Aviator SynWin is an updated and contextualized version of the synthetic work environment (SynWin; Elsmore, 1994) that requires participants to perform four separate, concurrent subtasks that each impose on different cognitive resources. We hypothesize that models derived using the LSTM architecture will accurately classify workload across the various difficulty conditions of the Aviator SynWin. We also expect that cerebral blood oxygenation levels will be negatively correlated with task performance and positively correlated with self-report workload ratings. The planned analyses will explore the relationship between task difficulty, performance, self-reported workload, and explicit (physiological) workload in both simple and complex tasks, promoting a deeper understanding of objective operator state and improving the potential for application of fNIRS to training settings.

METHOD

Participants

Seventeen participants (76% male) with a mean age of 27 (18-39, SD=5.16) were recruited from the Wright State University community area and comprised both students and professionals. All participants were at least 18 years old, native English speakers with normal or corrected-to-normal vision. None of the participants had a history of serious medical/mental illness and most (94%) engaged in at least 1-5 hours of video game play per week. Participants provided informed consent to participate and were compensated for their time.

Study Design

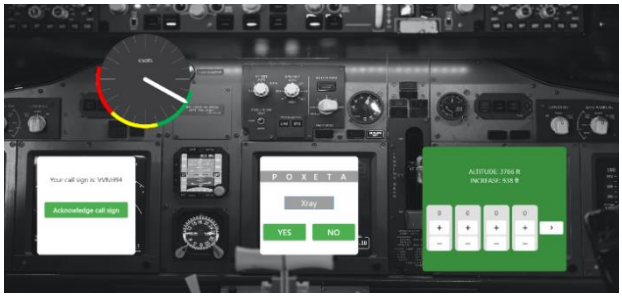

Figure 1. The Aviator SynWin environment is comprised of visual monitoring (upper left), auditory monitoring (lower left), memory recall (lower middle), and arithmetic tasks (lower right).

The present research study was designed to explore the calibration of workload models by utilizing a series of controlled experimental tasks to induce and measure changes in participant workload. Participants performed standard cognitive tasks intended to provide calibration data for model training. These included the n-back, the continuous performance test (CPT; Rosvold et al., 1956), and the Flanker task (Eriksen & Eriksen, 1974). Participants also performed an operationally relevant task, Aviator SynWin, that required them to simultaneously attend to a visual monitoring task, an auditory monitoring task, a memory recall task, and an arithmetic task (see Figure 1). Both the standard cognitive tasks and the Aviator SynWin task included multiple conditions that varied by level of task difficulty. The study employed a within-subjects design, ensuring that each participant experienced all manipulations of the experimental conditions. Each standard cognitive task consisted of four blocks per condition, while the Aviator SynWin task included a single 10-minute block per condition.

During the series of tasks, participants were outfitted with several physiological sensors so that physiological response to workload changes could be tracked. For this paper, we exclusively looked at workload tasks (n-back and Aviator SynWin). For the n-back task, participants completed 1-back, 2-back, and 3-back conditions which were labeled as ‘low’, ‘medium’, and ‘high,’ respectively. The Aviator SynWin task also had low, medium, and high

difficulty conditions. For this task, difficulty was manipulated by increasing the frequency of events within the auditory monitoring, visual monitoring, and memory recall subtasks and the timing parameters for each were established through pilot testing. The resulting physiological timeseries datasets enabled the development of individualized LSTM deep-learning models for workload classification, while behavioral data was used to compare conditions and tasks across subjects.

Measures and Equipment

N-Back d'
To capture performance and compare across conditions, signal detection theory was used to calculate d' from n-back data using hit and false alarm rates. While d' scores typically range between zero and two, the highest possible sensitivity (perfect performance) would yield a d' of about 4.65. A score of zero would reflect the poorest performance.

Aviator SynWin Composite
To measure performance across Aviator SynWin conditions with a variable number of events across the four subtasks, a composite scoring method was developed. For the audio and memory subtasks, hits and false alarms were used to calculate d'. For the arithmetic and visual monitoring tasks, a min-max z-score was calculated and multiplied by three to ensure the values approximated the same effective range as the d' scores. This was done to preclude any one task from having an outsized contribution to the composite. These scores were then added together and divided by four as shown in Equation 1. A composite score close to zero reflects poor performance and a composite score close to 3 reflects high performance.

$$SynWin\ Composite = \frac{d'_{Aud} + d'_{Mem} + 3(Z_{Math}) + 3(Z_{Vis})}{4} \quad (1)$$

NASA-TLX Composite
The NASA Task Load Index (NASA-TLX; Hart & Staveland, 1988) was administered as a subjective workload measure. Equal weights were used; composite scores were calculated by adding subscale values together and dividing by six. A composite score close to zero reflects low workload and a score close to 100 reflects high workload.

Cerebral Oxygenation
Initial calculations utilized raw light intensity values from fNIRS. A 5-minute resting baseline was captured, followed by the application of a 6th-order lowpass Butterworth filter with a cutoff frequency of 0.1 Hz. Subsequently, an average baseline value for each wavelength was computed for each participant. The same lowpass filter was then applied to data from each experimental condition. Next, changes in optical density were calculated from these filtered light intensity values by computing the negative logarithm of each filtered value divided by the respective baseline average. The Modified Beer-Lambert Law, incorporating wavelength-specific constants, differential pathlength factor (DPF), and distance, was applied to derive concentrations of oxygenated hemoglobin (HbO) and deoxygenated hemoglobin (HbR), as well as the Cerebral Oxygenation Index (COI). HbO serves as a proxy measure for neural activation; higher HbO values are expected to indicate increased neural activity.

Machine Learning (LSTM)
LSTMs were applied to our labeled multimodal timeseries data with the goal of classifying task difficulty level per participant. This is done by combining LSTM layers in a deep learning architecture with other layers (e.g., linear/dense layers) and outputting the probability values to make a prediction. This process includes using repeating layers of LSTMs to take in a context window of multimodal timeseries data from a participant and receive a predicted difficulty level. The model includes many parameters, often referred to as hyperparameters, that define the weights and bias values used to optimize the output. These hyperparameters can be tuned to give the most accurate predictions based on the training data utilized. Since we were training a separate model for each participant to account for individual differences (as opposed to developing a general model), models and their associated performance results were tracked using the open-source tool MLFlow, which aids in the development of management of machine learning code.

Equipment
Two fNIRS devices were used for simultaneous data collection; participants wore the Plux Pioneer on the right side of their forehead and Aptima, Inc.'s MAVERIC prototype on the left (see Figure 2). For heart rate data, participants

Figure 2. Participants wore two fNIRS sensors (MAVERIC, top left; Pioneer, top right) and one heart rate chest strap (bottom).

also wore a PolarH10 chest strap. Two separate laptops were used; one was the participant machine running the task and survey tools and one was the researcher machine running the data acquisition tools. Software used included Neuro-Behavioral Systems (NBS) Presentation Version 24.0 for the standard task batteries, Lab Streaming Layer to acquire and synchronize data, OpenSignals for Plux Pioneer data acquisition, and custom-developed software including the Aviator SynWin task, data acquisition tools, and survey tools. Participants used a standard keyboard and mouse for task inputs. The participant screen was positioned 29cm from the edge of the table and participants were seated in a chair positioned close to the edge of the table. For all standard cognitive task battery stimuli, the letter height was approximately 1.4cm on the screen. Overhead lights were turned off to reduce the risk of ambient light pollution in the fNIRS signals and a directional lamp was used, pointed away from the participant, to add dim lighting to the room.

Experimental Procedures

This study was approved by Aptima's Institutional Review Board (IRB) before beginning participant recruitment activities. Participants completed the study in two sessions that occurred no more than three days apart. Some participants completed both sessions in a single day with a minimum 30-minute break between sessions. The first session consisted of the standard cognitive tasks to be used for calibration of cognitive models (e.g., workload), and the second session consisted of the Aviator SynWin task.

To begin the first session, informed consent documents were signed, and participants were outfitted with the fNIRS devices and Polar H10 sensors. Participants then completed the demographics, resting baseline, and a head-motion calibration. Each participant completed the n-back (1, 2, and 3), the flanker task (congruent vs incongruent), and the CPT (vigilance vs sustained attention to response tasks). The n-back was always the first task completed, but the order of conditions within each task and order of ensuing tasks were pseudo-randomized. Digital instructions were provided at the beginning of each condition. After completing each condition of each task, participants filled out the NASA-TLX and a task rating survey. The total required time for the first session was approximately two hours.

For the second session, informed consent was reaffirmed, sensors were re-equipped, and participants again completed the resting baseline and head-motion calibration. Afterwards, they were provided digital instructions for the Aviator SynWin task that included text, pictures, a video tutorial, and five minutes of practice on the medium difficulty setting of the task. Each participant completed a single, 10-minute block for each of the three difficulty conditions (low, medium, and high). Conditions for the Aviator SynWin were also pseudo-randomized, and participants completed the NASA-TLX and task rating survey after each condition. The total required time for the second session was around an hour and a half.

RESULTS

N-back Behavioral Data

To ensure task difficulty manipulations were effective, we assessed n-back performance score data. The performance scores across conditions of the n-back, as expressed by d', generally matched previous literature and showed the expected trend (see Figure 3). Performance scores (d') decreased as task difficulty increased. Table 1 contains a full account of performance expressed within the signal detection framework. Means are presented for each metric across all conditions, with standard deviations denoted in parentheses.

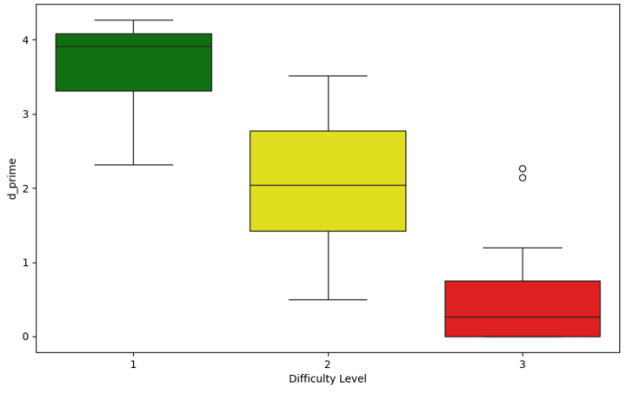

Figure 3. N-back performance data (d') across conditions

Table 1. Mean (and standard deviation) signal detection performance metrics for each n-back condition.

Condition	d'	Accuracy	Hits	False Alarms
1-back	3.67 (0.54)	99.11 (1.03)	38.18 (7.02)	1.44 (0.86)
2-back	2.05 (0.91)	93.39 (4.44)	34.65 (5.95)	7.68 (5.59)
3-back	0.54 (0.74)	82.29 (5.78)	24.36 (6.23)	19.98 (10.57)

Before comparing means across each of the difficulty conditions, we first assessed homogeneity of variance, sphericity and normality, as these are key assumptions for testing equality of means. Should any of these assumptions be violated, bias would be introduced into the test statistic. Levene's test ($F(2, 48) = 1.50$, $p = 0.24$) was used to assess the homogeneity of variances across conditions. Sphericity was assessed using Mauchly's test of sphericity ($W = 0.99$, $\chi^2(2) = 0.16$, $p = 0.92$). Neither test rejected the null hypothesis, indicating both assumptions were met. However, Shapiro-Wilk tests were conducted to assess the normality of each group. The 1-back ($p = 0.07$) and 2-back (p=0.83) conditions did not significantly deviate from normality, but the 3-back ($p = 0.0004$) showed a significant deviation from normality, thus violating the assumption.

Due to the violation of normality, we employed the non-parametric Friedman test, which does not have an assumption of normality, to compare the means. The Friedman test revealed a significant effect of condition on the d' scores, $\chi^2(2) = 32.12$, $p < 0.0001$. This result indicates that the means are not all equal across the conditions and performance was susceptible to changes in task difficulty. To further investigate which groups were different, post-hoc pairwise comparisons were conducted using the Wilcoxon signed-rank test with Holm's correction for multiple comparisons. The pairwise comparisons indicated significant differences between all pairs of n-back conditions. Specifically, there were significant differences between the 1-back and 2-back ($p < 0.001$, Hedges' $g = 2.12$), the 1-back and 3-back ($p < 0.001$, Hedges' $g = 4.75$), and the 2-back and 3-back ($p < 0.001$, Hedges' $g = 1.79$). The significant results from the Friedman test and subsequent post-hoc tests suggest substantial differences in the d' across the three conditions. The effect sizes, represented by Hedges' g, indicate strong differences between conditions, particularly between the 1-back and 3-back.

Aviator SynWin Behavioral Data

To ensure task difficulty manipulations were effective, we assessed Aviator SynWin performance data. Contrary to what was observed during pilot testing, the behavioral data for the Aviator SynWin task did not appear to follow the expected trend (Figure 4). Specifically, the overall composite scores (higher values correspond to better performance) show high variability and little separation in performance across conditions.

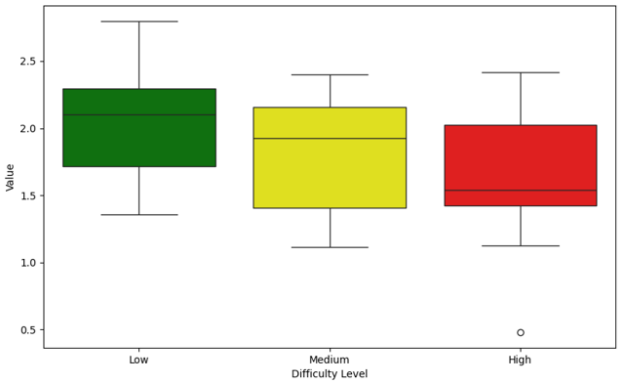

Figure 4. Aviator SynWin composite scores across conditions

We again assessed homogeneity of variance, sphericity and normality to ensure our test statistics would be appropriate. Levene's test ($F(2,48) = 0.36$, $p = 0.70$) was used to evaluate the homogeneity of variances across conditions. Sphericity was tested using Mauchly's test of sphericity ($W = 0.91$, $\chi^2(2) = 1.38$, $p = 0.50$). The normality of the data was assessed using Shapiro-Wilk tests for the low ($p = 0.65$), medium (p=0.22), and high (p=0.55) datasets, confirming data for each condition were normally distributed. The lack of significance in these assessments verified the assumptions for a repeated measures ANOVA.

Given the assumptions were met, a repeated measures ANOVA was performed to compare the means across the three conditions. The results revealed a significant effect of difficulty on the SynWin Composite Score, $F(2, 32) = 6.657$, $p = 0.0038$. This indicates that the means are not equal across the conditions and performance was susceptible to task difficulty manipulations. Pairwise comparisons were conducted to explore differences between conditions using the paired t-test with Bonferroni correction for multiple comparisons. The pairwise comparisons indicated a significant difference between the high and low difficulties ($p = 0.0061$, Hedges' $g = -0.78$). There were no significant differences between low and medium ($p = 0.056$, Hedges' $g = 0.54$) or high and medium ($p = 0.79$,

Hedges' g = -0.28). The significant results from the repeated measures ANOVA and subsequent pairwise comparisons highlight differences in the composite score between the low and high difficulties only.

Since the composite scores were not fully aligned with expectations, the team looked at individual subtasks. Examination of individual subtask data revealed potential ceiling effects in one subtask suggestive of failed task difficulty manipulation and a potential individual difference confound in another. While a full discussion of these data is outside the scope of the present paper, please see *Study Limitations* below for more details.

Self-Report Workload by Condition

We next assessed the impact of task difficulty on self-reported workload. When looking at NASA-TLX scores by task difficulty in both the n-back and the Aviator SynWin, the expected trend was observed. As seen in Figure 5, lower NASA-TLX ratings, indicating lower workload, were provided in the easier conditions (1-back and Low difficulty), and higher ratings were provided in the more difficult conditions (3-back and High difficulty). Note that due to technical issues in the acquisition software, there were three participants with partial NASA-TLX data. Specifically, two participants had partial n-back ratings data and one participant had partial Aviator SynWin ratings data.

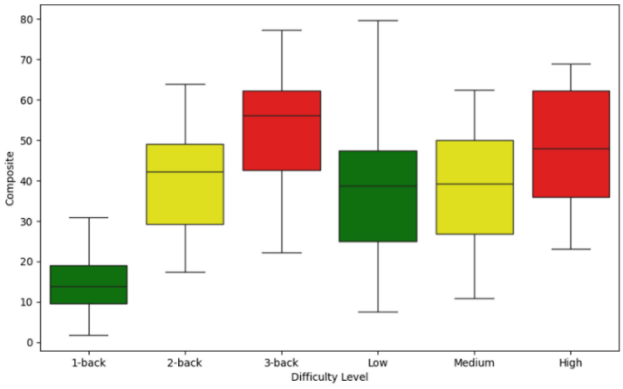

Figure 5. Average NASA-TLX composite across conditions for both n-back and Aviator SynWin tasks

Performance and Self-Report Workload

When looking at the overall relationship between performance scores across task conditions and self-report workload we observed significant negative correlations (see Table 2). There was a stronger effect for the n-back data (r = -0.72, p<0.0001), but the Aviator SynWin task data showed a similar significant negative correlation (r= -0.35, p=0.016) between self-reported workload and performance. These correlations are consistent with expectations, as task performance worsened when workload increased.

Cerebral Oxygenation

To look at the relationships between HbO and the other constructs of interest, HbO timeseries data were averaged for each condition. This resulted in six HbO values per participant: one HbO value per each of the three n-back conditions and each of the three Aviator SynWin conditions. These averages were used as an explicit measure of workload to test how the physiological data correlated with performance and with self-report workload for each task. As seen in Table 2, no significant relationships with HbO were observed in either the n-back or the Aviator SynWin.

Table 2. Correlation Matrix of Performance, Self-Report, and Physiological Data

	n-back d'	n-back NASA-TLX	SynWin Composite	SynWin NASA-TLX	HbO
n-back d'	1.00				
n-back NASA-TLX	-0.72***	1.00			
Aviator SynWin Composite	0.41**	-0.43**	1.00		
Aviator SynWin NASA-TLX	-0.25	0.28	-0.35*	1.00	
HbO	-0.05	0.14	-0.09	-0.24	1.00

$p < 0.05$*, $p < 0.01$**, $p < 0.001$***

Machine Learning (LSTM)

We tested our LSTM model using 14 participants' HbO timeseries data labeled by task condition (data from the remaining participants was unusable due to missing data). The tests included (1) using fNIRS data captured during

the n-back from Session 1 for training and validation, and using fNIRS data captured during the Session 2 Aviator SynWin for test data, and (2) using only fNIRS data captured during the Aviator SynWin for training and validation. This resulted in the training of 28 models and the results for both experiments can be seen in Figure 6. The accuracy used is a mean of the basic accuracy such that, for each sample, we get a score of 1 if the difficulty is predicted correctly, and 0 otherwise, for all windowed samples selected from the participants' data. This gives a maximum score of 1, and a minimum score of 0. Validation accuracy remained high for both experiments, meaning the model was able to learn the behavior of both Session 1 and Session 2 when applied to data coming from the same session. However, there was a large loss in accuracy when trying to generalize a model of workload based on HbO between sessions, as seen in the right plot of Figure 6.

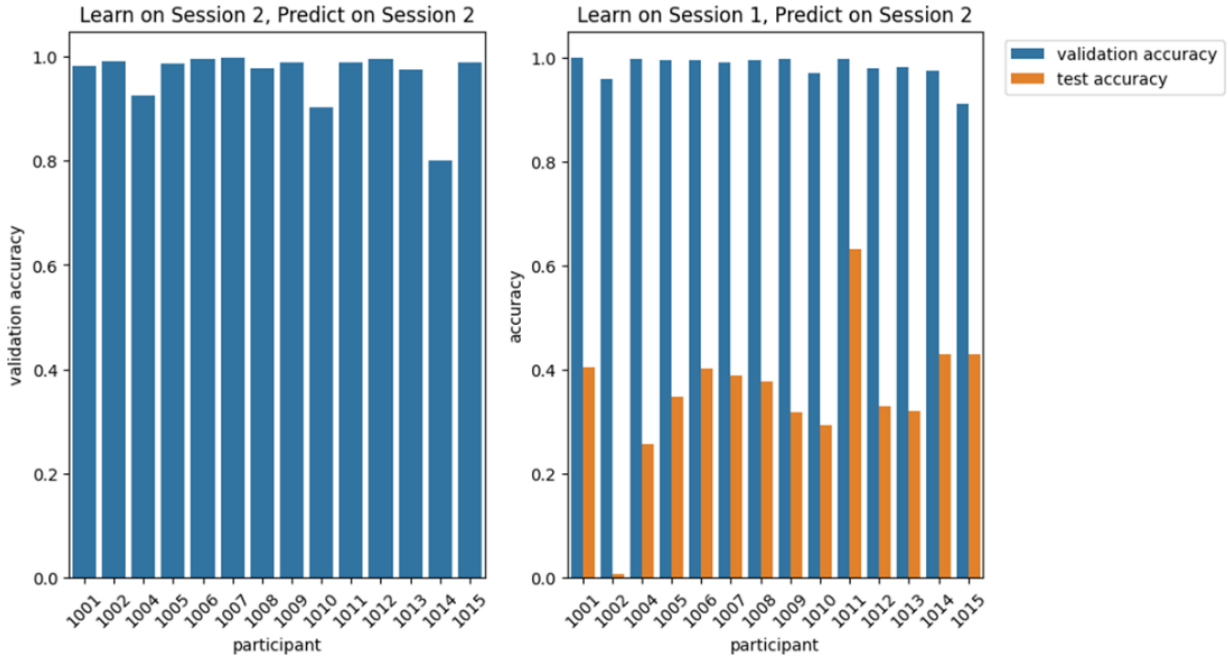

Figure 6. Accuracy results for Experiment 1 (left): trained on 80% session 2 data, validated on remaining 20% subset; and accuracy results for Experiment 2 (right): trained on 80% of Session 1 data, validated on remaining 20% subset, and tested on Session 2 data

DISCUSSION

This study explored the plausibility of transitioning fNIRS technology out of strictly laboratory environments and into training settings to unobtrusively provide access to robust state information that could be used to enhance training outcomes. We hypothesized that our performance data and self-report workload data would be consistent with observations from prior research. We also hypothesized that our LSTM approach would result in physiologically based workload models that could generalize across tasks.

In many ways, the results of our study align with prior literature. The performance data from the n-back task matched expectations with performance worsening as task difficulty increased. Further, our hypothesis that performance would be negatively correlated with self-report workload was supported. Both findings are consistent with prior research (e.g., Brouwer et al., 2012; Hancock & Matthews, 2019). With respect to the Aviator SynWin, the performance data only partially met expectations. While the Aviator SynWin task developed under this effort was based on a task utilized in previous research (e.g., Elsmore, 1994; Hambrick et al., 2010), this study represents a first attempt at implementing three levels of task difficulty. Therefore, task difficulty parameterization was developed through pilot testing and the results observed in the current study were not entirely consistent with the pilot data collected during task development. As expected, performance in the low and high conditions was

significantly different and overall, performance worsened as task difficulty increased. However, contrary to expectation, the performance scores from the medium difficulty condition were not significantly different from the other two conditions. The high variability observed in the scores within each condition suggests that our task parameterization potentially needs improvement to achieve separable performance across all conditions. Nevertheless, our hypothesis that Aviator SynWin performance would be negatively correlated with self-report workload was still supported.

With respect to the fNIRS data, the results indicate more work is needed to successfully capture explicit workload with low-cost, low channel fNIRS devices. We did not find support for our hypothesis that HbO would be negatively correlated with task performance and positively correlated with self-report workload ratings. This could be due to the smaller than desired sample size and the high variability in the dataset. It is also probable that our choice to average the neural data across the entire duration of each condition washed out meaningful trends that occurred within the timeseries. This notion is supported by the results of our modeling activities. Namely, our models performed well when validated on data from the same task with which they were trained, suggesting that the timeseries did contain meaningful patterns within each condition. We were successful in prototyping individualized algorithms capable of taking a historic window of a participant's physiological data and predicting the difficulty level within tasks. It is perhaps not surprising that our classification models suffered a loss in accuracy when attempting to generalize between the n-back and Aviator SynWin tasks given the mismatch in behavioral findings (i.e., presence/absence of a separable medium difficulty condition) and the physiological findings. Due to the preliminary nature of this work, further exploration is needed to determine if changing the model architecture allows behavior to be learned between tasks, or if the differences in the n-back and Aviator SynWin tasks are significant enough that the underlying timeseries data do not share the same signatures.

Study Limitations

The sample size of the present study was limited by access to the intended participant pool. Given the large variability in many of the measures, we likely would have benefited from a larger sample. The quality of our workload classifications was further limited by the parameterization of the Aviator SynWin difficulty levels. Despite pilot testing, we did not see the expected performance variation across the low, medium, and high difficulty conditions, making it more difficult for our LSTMs to make appropriate classifications. Finally, the classification accuracy results presented reflect the team's first attempt at model generation. The team unfortunately did not have sufficient opportunity for hyperparameter tuning which could possibly improve accuracy in the future.

Conclusions and Future Work

While not all hypotheses were fully supported, the findings of the present study provide enough support to warrant additional investigation. In the future, the team would like to collect additional data with a similar protocol to improve statistical power. Future data collection opportunities might also include the incorporation of additional tasks to aid in creating a more generalizable workload model capable of reliably classifying workload across a variety of operational contexts. With respect to the Aviator SynWin task specifically, a series of small pilot studies could also be conducted to improve the parameterization of the task difficulty conditions. Additional testing would allow us to fine-tune subtask event timing so that the performance data across the conditions is more distinct and separable to aid the machine learning efforts. During these pilots, a general mathematical skills assessment should be added to the self-report battery so that individual differences in mathematical ability can be controlled for in task performance analyses.

Additionally, there are alternative data exploration techniques that could be pursued with the existing dataset. For example, it could be beneficial to account for task performance when generating models given previous findings that task performance moderates the relationship between task demands and neural activation in the PFC (Meidenbauer et al, 2020). The team could also explore additional analytic methods to characterize the relationships between the physiological timeseries data, behavioral data, and self-report data (i.e., beyond the machine learning work completed to date) such as using various windowed averages. Finally, there are portions of the existing dataset that remain unexplored, including the CPT and Flanker task data, which could be used to develop additional state classification models. Taken together, these activities would help guide future model development by answering the question of whether it is necessary to change the model architecture.

The test and evaluation activities outlined above resulted in a large data set that allowed us to explore the utility of low cost fNIRS systems and a novel multitask environment, Aviator SynWin. The experimental outcomes and analytics led to the generation of a machine learning pipeline that synchronizes multimodal data, applies preprocessing, and produces individualized models of workload classification. While these models could benefit from additional development, they offer a promising start towards a framework for operational state assessment using unobtrusive, neural sensors that could prove beneficial for military training use cases.

ACKNOWLEDGEMENTS

This work was supported by the Defense Health Agency and U.S. Navy under Contract No. W81XWH21C0083. The views, opinions, and/or findings contained in this paper are those of the authors and should not be construed as an official position, policy or decision of the Defense Health Agency or the U.S. Navy.

REFERENCES

Brouwer, A. M., Hogervorst, M. A., Van Erp, J. B., Heffelaar, T., Zimmerman, P. H., & Oostenveld, R. (2012). Estimating workload using EEG spectral power and ERPs in the n-back task. *Journal of neural engineering*, *9*(4), 045008.

Berka, C., Levendowski, D. J., Lumicao, M. N., Yau, A., Davis, G., Zivkovic, V. T., Olmstead, R.E., Tremoulet, P.D., & Craven, P. L. (2007). EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviation, space, and environmental medicine*, *78*(5), B231-B244.

Elsmore, T. F. (1994). SYNWORK1: A PC-based tool for assessment of performance in a simulated work environment. *Behavior Research Methods, Instruments, & Computers, 26(4),* 421-426.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & psychophysics*, *16*(1), 143-149.

Fuster, Joaquin. (2014). The Prefrontal Cortex Makes the Brain a Preadaptive System. Proceedings of the IEEE. 102. 417-426.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in Psychology, 52*, 139-183.

Hambrick, D. Z., Oswald, F. L., Darowski, E. S., Rench, T. A., & Brou, R. (2010). Predictors of multitasking performance in a synthetic work paradigm. *Applied cognitive psychology*, *24*(8), 1149-1167.

Hancock, P. A., & Matthews, G. (2019). Workload and performance: Associations, insensitivities, and dissociations. *Human factors*, *61*(3), 374-392.

Herff, C., Heger, D., Fortmann, O., Hennrich, J., Putze, F., & Schultz, T. (2014). Mental workload during n-back task—quantified in the prefrontal cortex using fNIRS. *Frontiers in human neuroscience*, *7*, 935.

Kirchner, W. K. (1958). Age differences in short term retention of rapidly changing information. Journal of Experimental Psychology, 55, 352-358.

Meidenbauer, K. L., Choe, K. W., Cardenas-Iniguez, C., Huppert, T. J., & Berman, M. G. (2021). Load-dependent relationships between frontal fNIRS activity and performance: A data-driven PLS approach. *NeuroImage*, *230*, 117795.

Ranchet, M., Morgan, J. C., Akinwuntan, A. E., & Devos, H. (2017). Cognitive workload across the spectrum of cognitive impairments: A systematic review of physiological measures. *Neuroscience & Biobehavioral Reviews, 80*, 516-537.

Rosvold, H. E., Mirsky, A. F., Sarason, I., Bransome Jr, E. D., & Beck, L. H. (1956). A continuous performance test of brain damage. *Journal of consulting psychology*, *20*(5), 343.

Pinti, P., Tachtsidis, I., Hamilton, A., Hirsch, J., Aichelburg, C., Gilbert, S., & Burgess, P. W. (2020). The present and future use of functional near-infrared spectroscopy (fNIRS) for cognitive neuroscience. *Annals of the New York Academy of Sciences*, *1464*(1), 5–29. https://doi.org/10.1111/nyas.13948

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive Science, 12, 257-28

Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation.