Day-to-Day Variability in Hybrid, Passive Brain-Computer Interfaces: Comparing Two Studies Assessing Cognitive Workload*

Samantha L. Klosterman, Justin R. Estepp, Member, IEEE, Jason W. Monnin, and James C. Christensen

Abstract—As hybrid, passive brain-computer interface systems become more advanced, it is important to grow our understanding of how to produce generalizable pattern classifiers of physiological data. One of the most difficult problems in applying machine learning algorithms to these data types is nonstationarity, which can evolve over the course of hours and days, and is more susceptible to changes resulting from complex cognitive function in comparison to simple, stimulus-based processes. This nonstationarity, referenced as day-to-day variability, results in the inability of many learning algorithms to generalize to new data. In previous work, we have shown that increasing the number of unique testing sessions used to form a learning set can improve the accuracy of classifying mental workload in a binary state paradigm. While this result was very promising, we did not address whether the additional discriminability was the result of a larger learning set or the uniqueness contributed by the testing sessions being spread over multiple days. Further, the simulation task used in this prior analysis was low-fidelity with respect to the task it attempted to model; whether these methods extend to more realistic task simulation environments has not been comparatively investigated. In this work, we compare these previous results to a second study, with a similar multi-day paradigm, that required participants to perform a more realistic simulation task. Comparative analysis of these two studies reveals that the improved generalization of the multi-day learning set is attributable, in large part, to the uniqueness of the multi-day paradigm. Further, this multi-day effect was also observed in the higher fidelity simulation study. These results help to validate the use of the multi-day learning set approach for improving overall system classification accuracy. Future studies should consider the use of multi-day designs for improving generalizability over other interesting dimensions.

I. INTRODUCTION

The application of neurophysiological data to braincomputer interfaces (BCI) and brain-machine interfaces (BMI), in particular to neurally-derived signals such as electroencephalography (EEG) [1]-[2], electrocorticography (ECoG) [4]-[5], and multi-unit neural recordings [6]-[7], has resulted in a number of ground-breaking research applications (for a series of excellent reviews, see [8]-[13]).

While each of these methodologies exhibit different strengths and weaknesses, we concur with the view that each has value and all should be the focus of continued, parallel research efforts [14]. In our particular focus on passive brain-computer interfaces (pBCI) [15] and hybrid brain-computer interfaces (hBCI) [16]-[18], EEG is, currently, the most popular approach, largely due to its relative non-invasiveness. Given its popularity, there has been extensive prior work addressing the signal processing of [14] and application of machine learning to [19] EEG. Furthermore, the advancement of EEG recording technology over the last 10 years [20]-[21] gives the promise of an even more elegant solution to use outside of the laboratory environment.

There are, however, several important issues in using EEG as an input to a BCI that need to be resolved [14]. One particular issue that we have focused on is the nonstationarity of EEG over the course of time. While an individual’s EEG response to simple lab tasks has been shown to be stable over time [22]-[25], building a learning set from multiple data collection sessions has been shown to increase the performance of machine learning algorithms based on event-related potential (ERP) components [26]. This result suggests that, in an ERP paradigm, unique features of the ERP that may evolve over time can be effectively utilized.

In our own work using spontaneous, spectral features of EEG to assess cognitive states, such as workload, in complex task simulations [27]-[29], we have shown a similar improvement in learning algorithm accuracy on independent test sets when multiple sessions, collected over the course of hours and days, are added to the learning set [30]. Two shortcomings within our work here are that it was not determined whether the improvement in discriminability was from the larger learning set or the unique data obtained using multi-day learning set and whether these methods extend to a more realistic task environment. While we demonstrated these results using several machine learning algorithms, the most successful approach involved using an Artificial Neural Network (ANN) and treating both the learning and test sets as samples from different distributions (computing separate distribution statistics for the learning and test sets prior to converting the feature vectors to z-scores). While this result in and of itself suggests that varied parameterization of spectral feature distributions may account for decreased learning algorithm performance, it also presents a problem in that the sample distribution statistics of the test set may not be well-characterized a priori in a real-time system.

In this present work we examined whether the increased test set classification accuracy seen in a previous study [30] was simply the result of a larger learning corpus or whether there was a unique contribution to improved generalization

*With appreciation, we acknowledge that this work is supported by the Air Force Office of Scientific Research (AFOSR) under Air Force Research Laboratory Task 15RHCOR234 and the 711th Human Performance Wing Chief Scientist Seeding Program.

S.L. Klosterman and J.W. Monnin are with Ball Aerospace & Technologies Corp., Fairborn, OH 45324-6222 USA (corresponding author phone: 937-938-3604; email: [skloster, jmonnin]@ball.com).

J.R. Estepp and J.C. Christensen are with the 711th Human Performance Wing, Air Force Research Laboratory, Wright-Patterson Air Force Base, OH 45433-7951 USA (e-mail: [justin.estepp, james.christensen.7]@us.af.mil).
due to sampling the learning set from a multi-day data collection paradigm. Also, in an analysis of a second dataset [31], we will apply these same techniques to classifying mental workload in a more complex, realistic task simulation environment. In order to simulate a real-time p/hBCI system, this analysis will be accomplished by implementing a real-time classification system (using distribution statistics from the learning set to z-score the test set) post-hoc using data that was collected from open-loop portions of the studies [30]-[31].

II. METHODS

A. Data Set I: MATLAB (Low-Fidelity Task Environment)

The dataset collected from the low-fidelity task environment is fully described in our previous work [30]. In brief, 8 participants (3 male, mean age of 21.1 years) completed an approximately 3-hour long task training session (over the course of one or two days) until performance parameters attained asymptote with minimal errors. Training was followed by three data collection sessions per day for five days spread over the course of a month. The task environment (Fig. 1) was the Multi-Attribute Task Battery (MATB; [32]). In order to more closely represent the operation of remotely piloted aircraft (RPA), the compensatory tracking task was fully automated, but the monitoring (lights and dials), communication, and resource management subtasks were each simultaneously presented to the participant. In each session, 5 minutes of both low and high task demand were presented (in random order). During task performance, nineteen channels of EEG according to the International 10-20 system [34] were recorded, referenced to a single mastoid. Horizontal and vertical electrooculogram (HEOG and VEOG, respectively) were recorded by placing electrodes on the outer canthus of each eye (for HEOG) and above and below the left eye (for VEOG). Additionally, a bipolar electrocardiogram (ECG) signal was recorded. All electrodes were standard 9 [mm] tin cup electrodes.

B. Data Set II: RPA (High-Fidelity Task Environment)

The second dataset was collected from a “suppression of enemy air defense” simulation for controlling multiple RPAs [33] (Fig. 2). A total of 10 participants (8 male, mean age of 23.0 years) completed an approximately 10-hour long task training protocol (over the course of several days) designed to achieve asymptotic task performance and mitigate any potential task learning effects. Task difficulty was manipulated by requiring the participants to control either 2 groups of 4 RPAs (8 RPAs total, ‘2-Fly’) in the low task difficulty condition or 4 groups of 4 RPAs (16 RPAs total, ‘4-Fly’) in the high task difficulty condition. Without direct interaction, the RPAs autonomously flew with a northern heading over a course of pre-defined waypoints, but their flight path could be interactively redirected with the restriction that the RPA could not back-track to a previously visited waypoint. During flight the participants were required to navigate the RPAs to densely populated areas of interest (AOI). For each AOI, the participant was required to first download a ground image of the AOI and perform a target identification task. Based on the target type the participant was then required to make an appropriate weapons pairing to successfully prosecute the target. Depending on the number and type of targets present in any ground image, multiple RPAs could be required to successfully prosecute the ground vehicles, as a single RPA did not carry all munitions types (or, it could have been the case that required munitions had been previously exhausted on a prior target). Successful prosecution of a ground image required perfect accuracy in both target identification and weapons pairing.

In a single session (taking place during one day), the participants completed a single 2-Fly (low task difficulty) trial and a single 4-Fly (high task difficulty) trial. Each of these trials lasted approximately fifteen minutes, although the time for each trial could vary by several minutes depending on how the participant chose to re-direct the RPAs during flight. In addition to these two trials, a third, mixed workload trial was performed (hereby called the ‘mixed’ trial). The mixed trial began with a fifteen minute 4-Fly, followed immediately by a five minute 2-Fly, and ended with a second fifteen minute 4-Fly. As with the 2-Fly and 4-Fly single workload trials, the exact duration of the mixed trial could vary depending on how the participant chose to re-direct the RPAs during flight.

During the RPA simulation task, 5 channels of EEG (Fz, F7, Pz, T5 and O2) from the International 10-20 system [34] were recorded, along with HEOG, VEOG and ECG as

![Figure 1. User interface for the Multi-Attribute Task Battery (MATB) task. The four subtasks (System Monitoring, Tracking, Communications, and Resource Management) are shown in the left center columns on the interface. The right column shows Scheduling and Pump Status information windows. The Scheduling and Tracking tasks were disabled for this study.](image1)

![Figure 2. User interface for multi-remotely piloted aircraft (RPA) control task. The wedge shaped symbols at the bottom represent RPAs that the operator is controlling; the missile-shaped symbols represent air defense sites that need to be engaged.](image2)
previously described. These data channels were recorded with a BioRadio 110 wireless telemetry system (Great Lakes Neurotechnologies, Inc., formerly Cleveland Medical Devices, Inc., Cleveland, OH, USA), sampled at 200 [Hz], and natively bandpass filtered between approximately 0.4 [Hz] and 53 [Hz]. All electrodes were standard 9 [mm] tin cup electrodes.

C. Data Preprocessing

For both datasets, all EEG channels were corrected of HEOG and VEOG artifact, although a post-hoc regression method was used for the MATB dataset, and an on-line adaptive filter method was used for the RPA dataset [35]-[36]. Inter-beat interval (IBI), as the time between successive R-waves, was computed for each ECG time series using a real-time algorithm [37]. A real-time blink detection algorithm [38] was used to calculate inter-blink interval (IBLI) in [ms] for successive eye blinks as detected in VEOG.

Spectral band power (over a 10 [s] window, using a 10 [s] Hanning window and either the fast-Fourier transform or discrete Fourier transform as appropriate with respect to sampling rate) in the traditional clinical EEG bands was also calculated for both datasets (for both the EEG, HEOG and VEOG channels). These bands were defined as delta (0.5–3 [Hz]), theta (4–7 [Hz]), alpha (8–12 [Hz]), beta (13–30 [Hz]) and gamma (31–42 [Hz]).

D. Feature Sets

For both datasets, a ten second window was used to average IBI. Blink rate (in blinks per minute) was derived from IBLI over a 30 [s] window. For each window, a single feature vector was created; successive feature vectors were created by advancing the window by one second and re-calculating the features, thus resulting in a feature update rate of 1 [Hz]. This resulted in 37 features: 5 features for each of the 5 EEG, HEOG and VEOG channels, plus IBI and blink rate.

In regard to window length, we have previously shown that longer windows, such as the forty second window used previously with the MATB dataset, produce more accurate classification results [39], a likely effect of increased signal-to-noise ratio in the feature data resultant from a larger data sample. In the work presented here, we attempt to make direct comparisons of classification accuracy from the two unique studies by reprocessing the MATB dataset so that it is analogous to the RPA dataset for which some data were processed on-line using different parameters in a closed-loop paradigm. Although that specific closed-loop data is not analyzed here, this allows the MATB results to be reasonably compared to what would have been obtained from real-time classification of the open-loop portions of the RPA study. To achieve this, we (1) used the same features (37), (2) the same averaging window (10 [s]), and (3) sampled the learning and test sets from the same parameterized distributions, which uses mean and standard deviation calculations form the learning set to z-score the test set in the real-time system implementation.

E. Machine Learning Algorithms

For both datasets, three machine learning algorithms are used: the ANN, a kernel-based support vector machine (SVM) with a linear kernel (LIN-SVM), and linear discriminant analysis (LDA). The ANN is a three-layer feed-forward network with the backpropagation algorithm implemented for learning [40]-[41]. For the ANN, a separate validation set (25% subsampled from the learning set) was used to mitigate overfitting. For each of the successive learning iterations, the learned weights and biases were updated only if the feed-forward error on the validation set was equal to or less than the error obtained from the previous learning iteration. After convergence (meeting a pre-defined error threshold), the weight and biases were fixed and used in the feed-forward direction to derive performance accuracy on the test sets. The LIN-SVM was constructed using the kernel-based least-squares SVM (LS-SVMlab v1.8 MATLAB toolbox [42]). Given that construction of the SVM does not adhere to an iterative stopping rule as with the ANN, the validation set was re-added to the learning set. The LDA was implemented via the MATLAB Statistics Toolbox (MATLAB R2010a, Statistics Toolbox Version 6.2, The Mathworks, Natick, MA, USA). The same learning and test sets as used with the SVM were also used by the LDA.

For each learning algorithm, a k-fold (nested, for the ANN) cross-validation (k = 10) was used to estimate an unbiased accuracy on the test sets. Classification accuracies are presented as the average accuracies across the 10 folds for each participant, with each participant contributing an N = 1 sample to the accuracy distributions (N = 8 for the MATB dataset; N = 10 for the RPA dataset).

F. Day Combinations

To investigate the contribution of adding additional days for training the machine learning algorithm, we incrementally added n days to the learning set, beginning with n = 1, until the number of days in the learning set was n = M-1, where M was the total number of days available (M = 5 for the MATB dataset; M = 3 for the RPA dataset), given that including all days are in the learning set when n = M would result in an empty test set (Fig. 3). Day combinations for the learning set were formed using a full permutation of the possible combinations; i.e., for the n = M-1 case for the MATB dataset, there were 5 possible permutations, each of which resulted in one of the five days being reserved for the test set. Since within-day accuracies (data randomly subsampled from a day and withheld as a test set) are typically at ceiling (likely due to the less restrictive assumption of independence resulting from the window overlap) and we are primarily interested in improving classification accuracy across days, we only report test set classification accuracies for days not included in the learning set. Previously, we have shown that there is not a significant difference in test set accuracy due to the amount of time elapsed between learning and test sets when time elapsed is greater or equal to one day [43]; therefore, we collapse test set accuracy across all days not included in the learning set.
The objectives of this research were to (1) determine if the improvement of classification accuracy seen in the MATB dataset analysis was a result of increasing the number of learning examples or by having a longer time course over which these learning examples were recorded, and (2) to see if these methods extend to a higher-fidelity, more realistic...
task environment. Looking at the low-fidelity task classification accuracies using the full dataset for training, the results are quite different than previous results [25]. The present results show that the ANNs consistently produced the lowest classification accuracies of the three classifiers where previous results showed the ANN outperforming the other classifiers (LDA and LIN-SVM). This could be a result of decreasing the number of features from 193 to 37 and treating the learning and test sets as samples from the same parameterized distribution. The reduction in the number of features may have significantly improved classification accuracy for the linear classifiers (which are known to be less susceptible to overfitting), thereby allowing them to outperform the ANN [46]. The overall classification accuracies from the current results are also not as successful as previous reported and could be a result of both the reduction of features and the different windowing methods used [39]. Previous results used a 40-second window with 35-second overlap. However, even considering these differences in data preprocessing methods, there is still a significant effect of days seen in the current results which supports the idea that by adding days to the training set an improvement in classification accuracy is achieved.

The subsampled datasets produced slightly lower classification accuracies as compared to using the full dataset (the main effect of subsampling was significant). However, the addition of days to the subsampled learning set significantly improved the overall test set accuracies (i.e., the main effect of days was significant). The interaction effects between sample size and days in training set can be easily interpreted from the data in Fig. 4; the combined effect of full samples and the most number of available days in the training set resulted in the highest overall classification accuracies.

While none of the main effects for classifier, day, and sample were strictly significant, for the RPA dataset, the main effects for both day and sample did approach significance at p = 0.165 and p = 0.141, respectively. Because the RPA study was limited to only 3 total data collection days, the maximum number of days that could be present in the learning set while leaving at least one day to compose the independent test was 2; however, comparing the 1- and 2-day combinations in Fig. 4 and Fig 5, the results from these learning set combinations in the MATB study are very similar to those seen in the RPA study, an observation that strongly suggests similarly significant effects could have been obtained from the RPA study had it utilized more than 3 days in its test paradigm.

From the totality of these results it can be concluded that discriminability between high and low workload levels is equally achievable in both low and high-fidelity task simulation environments. Improvement in classification accuracy using the multi-day paradigm is not only seen in a simple lab task but was also observed in the more realistic RPA simulation, with the speculative caveat that the effects approaching significance in the RPA study may have been limited because the protocol ended after three days of testing. These results also support the idea that the increase in classification accuracy by adding days to the training set has a unique outcome due to the longer time course over which these data were sampled, although there appears to be an effect of the total size of the learning corpus that is appreciable, as well. Based on these results, “a practical solution to reliable … classification could be to conduct brief recalibration sessions each day following a longer training session” [30]. Future improvements could include training a classifier with a learning set that contains small subsets of data from multiple days (M > 5) collected over the course of weeks, months, or years.

The multi-day learning approach in p/hBCI systems appears to be a valid and promising construct to pursue for achieving consistently generalizable learning accuracies when testing on independent test sets. Further testing of this method will, of course, warrant additional study. After four days of data in the learning set for the MATB study, the classifier accuracies do not yet appear to be reaching a

![Figure 4. Average classification accuracy across subjects represented using the proportion of epochs correct for the MATB dataset (low-fidelity task). The average accuracies are displayed for each classifier and the number of days included in the training dataset, full or subsampled. Error bars are standard errors.](image1)

![Figure 5. Average classification accuracy across subjects represented using the proportion of epochs correct for the RPA dataset (high-fidelity task). The average accuracies are displayed for each classifier and the number of days included in the training dataset, full or subsampled. Error bars are standard errors.](image2)
plateau; further improvements in classifier accuracy may be obtained with protocols that make use of more than 5 days of testing (where one day must be allocated to the independent test set). These results also indicate that these methods may be suitable for even more realistic and relevant tasks; future studies should seek to advance realism by having participants perform very relevant and necessary tasks and duties rather than carefully controlled laboratory simulations. As generalization improves temporally as more unique days’ data are added to the learning set, it may also be interesting to study other dimensions of variability over which generalization has historically been problematic, namely varying individuals and tasks with similar cognitive paradigms.

ACKNOWLEDGMENT

The authors would like to extend their sincerest gratitude to Iris E. Davis and Margaret A. Bowers (Ball Aerospace & Technologies Corp.) for their assistance with preparation for these studies and data collection.

REFERENCES


