A Neurophysiological Sensor Suite for Real-Time Prediction of Pilot Workload in Operational Settings

Trevor Grant[1(✉)], Kaunil Dhruv[1], Lucca Eloy[1], Lucas Hayne[1], Kevin Durkee[2], and Leanne Hirshfield[1]

[1] Institute of Cognitive Science, University of Colorado, Boulder, CO, USA
trgr2496@colorado.edu
[2] APTIMA, Inc., Dayton, OH, USA

Abstract. In recent years, research involving the use of neurophysiological sensor streams to quantitatively measure and predict the level of mental workload experienced by an individual user has gained momentum as the complexity of the tasks operators have experienced in heavily computerized contexts has continued to expand. Despite the promising results from many empirical studies reporting successful classification of workload using neurophysiological sensor data, accurate classification of workload in real-time remains a largely unsolved problem. This research aims to both introduce and examine the efficacy of a new research tool: Tools for Object Measurement and Evaluation (TOME). The TOME system is a toolset for collating and examining neurophysiological data in real time. Following a presentation of the system, and the problems the system may help to solve, a validation study using the TOME system is presented.

Keywords: Mental workload · Physiological sensors · Data acquisition

1 Introduction

In recent years, research involving the use of neurophysiological sensor streams to quantitatively measure and predict workload experienced by an individual has gained momentum with the complexity of its applications ranging from driving cars [1] to playing music [2] and web surfing [3]. Such systems often pair a neurophysiological measurement modality such as functional near-infrared spectroscopy (fNIRS) or electroencephalogram (EEG) with other physiological sensors such as electrocardiogram (ECG), electrooculogram (EOG), respiration rate sensors and galvanic skin response (GSR). Data collected from these modalities are then fused together to build classifiers trained to discretely predict mental states from these physiological signals using machine learning techniques [4]. While these studies effectively correlate performance in simulated tasks with workload, their extension to a practical setting is often limited by the footprint and portability of these sensors. This is especially true in the aviation domain, where pilots are often wearing helmets or other flight gear and must maneuver aircraft in

C. Stephanidis et al. (Eds.): HCII 2020, LNCS 12425, pp. 60–77, 2020.
https://doi.org/10.1007/978-3-030-60128-7_5

both simulated and actual flight environments. Naturalistic environments such as these necessitate the use of wearable sensor suites that are highly practical for deployment in operational settings. There are often tradeoffs, however, between a sensor's physical profile, portability, and efficacy with references to recording a reliable signal which may be used for both analysis and predictive modeling. A minimal yet comprehensive sensor suite that allows a highly practical and efficient data collection procedure is required to build robust models suited for in-flight studies.

Despite the promising results from many empirical studies reporting successful classification of workload using neurophysiological sensor data, the ability to accurately classify the level of user workload in real-time remains a largely unsolved problem. This issue arises because models that were trained on small laboratory datasets often fail to generalize beyond the original dataset. These models fail to transfer to new sensors, new contexts, new people (or even to the same person in a different day). Neurophysiological datasets are high dimensional in nature, and they should be trained on a suitable number of instances to enable the creation of generalizable models. Several recent papers have begun to detail these challenges [5–7].

With these goals in mind, this research makes the following three contributions to the research domain: First, we present the Tools for Objective Measurement and Evaluation (TOME) system, a diagnostic tool-set for cost-effectively supporting test and evaluation practitioners using a highly practical suite of neurophysiological sensors. With the suite of sensors, TOME supports real-time secure cloud-based data acquisition and data storage via an easy to use graphical user interface (GUI). The TOME system works with a suite of neurophysiological sensors that were validated and selected based on their ability to maintain cost effectiveness, portability, comfort and practicality for use in ecological flight simulation scenarios, while still maintaining quality of the psychophysiological data. These sensors (detailed later in this paper) include functional near-infrared spectroscopy (fNIRS), Electrocardiogram (ECG), Electrodermal Activity (EDA), respiration, and eye tracking sensors. Second, we present the results of an empirical study where difficulty levels were manipulated while participants piloted an F-18 Aircraft in an X-Plane flight simulator environment. The TOME system was used to collect neurophysiological data in 10 participants and a support vector machine (SVM) was trained on the resulting data to predict participant workload caused by the changes in difficulty. Third, we expand upon our predictive analyses to include an evaluation of the value and sensitivity of the different data streams in the overall classification accuracy of operator workload.

1.1 Background and Literature Review

Defining and Measuring Mental Workload: A dearth of research has explored the construct of mental workload, and while there remains contention regarding the exact definition, most researchers agree that mental workload is a product of the demands of a task and the mental capacity of the person performing the task [8]. Techniques for measuring mental workload can be divided into subjective ratings, secondary-task behavioral measures, and physiological measures [9]. Common subjective measurements such as the NASA-TLX [10] and the Bedford Workload Scale [11] have been widely used to

assess workload, due in part to the ease by which they can be administered. These surveys have been widely recognized to be sensitive to mental workload, but they suffer from the same drawbacks of other self-report surveys, including the inability of people to accurately self-assess their own changing workload as well as the fact that they are administered after a task has been completed, which lacks real-time information [8, 9, 12]. Secondary task performance is another common way to assess mental workload. This technique assumes that decrements in secondary task performance are due to the combined task load exceeding a person's workload capacity. While this technique does not suffer from the subjective issues of self-report scales, some researchers have criticized this technique as dual task decrements in performance vary with different allocation of resources [9, 13, 14], allowing for researchers to only infer workload from performance.

To overcome the drawbacks of self-report and secondary task performance measures for assessing workload, researchers have turned directly to cognitive and physiological sensors in order to acquire real-time, objective measures of workload. To this end, recent research has used a myriad of sensors to measure and predict workload. This includes brain measurement with EEG [15, 16] and fNIRS [12, 17], or physiological measurements such as heartrate, galvanic skin response, respiration rate, or pupil diameter [18, 19]. In order to gain a more complete picture of workload, researchers have begun to merge data streams from various neurophysiological devices, as the data are often complementary [4, 20]. For example, Molina et al. were able to classify four levels of mental workload by combining different signals including EDA, electrocardiogram, photoplethysmography (PPG), EEG, temperature and pupil dilation during a web browsing task [3]. See Lohani [21] for a thorough review of recent empirical psychophysiology research focused on measuring workload and other related constructs (e.g., attention), which includes details about the biological mechanics underlying psychophysiological sensing (heartrate, galvanic skin response, respiration, pupil size, etc.) as well as relative strengths and limitations of each measure.

Predicting Mental Workload in Driving and Flight Settings: As noted above, our research is focused on prediction of pilot's workload changes during flight, which dovetails with research in the driving domain, as both tasks (flying a plane and driving a car) require continuous attention and multitasking to maintain performance. Several researchers have explored mental workload prediction during driving [21–24]. For example, recent work explored the combination of ECG, EOG, EEG and/or fNIRS modalities to investigate the effect of sleep deprivation on a subjects' performance in a simulated car driving study [22]. Though aircraft pilots face less traffic compared to their on-ground counterparts, piloting is a complex multi-tasking activity that requires both skill and technical expertise [25]. From an HCI perspective, piloting an aircraft is a cognitively demanding, resource intensive task, exercising working memory to satisfy task demands [26]. Numerous studies have used neuro-imaging modalities such as EEG [19, 27–29] and fNIRS [30] paired with other physiological sensors to evaluate neuro-physiological correlates of pilots cognitive workload in a simulated flight environment.

Although machine learning has been used to predict mental workload on cognitive and/or physiological sensor data in many prior empirical studies with promising accuracies, transitioning those successes outside of the laboratory remains challenging. This

prior research suffers from two challenges: First, data collection is a time consuming and laborious task, with prior models often trained on very small datasets that fail to transfer to other domains, other participants, other sensor manufacturers. Second, it can be difficult to compare and collate results from studies as data sharing is not common and results differ by sensors used, number of participants in study, number of workload states predicted, operationalization of 'ground truth' workload per study, as well as whether or not models are built within subjects or across subjects. Therefore, despite the advancements reported from many empirical studies reporting successful classification of workload, a number of recent papers have shed light on the above issues [5–7].

1.2 The TOME System

TOME, The Tools for Objective Measurement and Evaluation system, is a diagnostic toolset for cost-effectively supporting test and evaluation practitioners and augmenting system acquisition decisions through advanced workload measurement and performance assessment strategies. TOME's current implementation includes the following five psychophysiological sensors, which were selected for inclusion based on their ability to maintain cost effectiveness, portability, comfort and practicality for use in ecological flight simulation scenarios, while still maintaining quality of the collected data 1) a 2-Channel functional near-infrared spectroscopy (fNIRS) device from PLUX for measuring blood flow in the frontal cortex, 2) a Zephyr Bioharness for measuring respiration, 3) a Polar HR10 shirt for measuring heart rate, 4) a Polar Smartwatch for measuring galvanic skin response data, and 5) a desk-mounted Tobii eyetracker for assessing pupil diameter (Fig. 1).

Fig. 1. A screenshot of the TOME dashboard.

The Body Area Network (BAN) transmitter is a service application that currently runs on Android mobile devices. The BAN transmitter's main responsibility is connecting to the various sensors a person is wearing and transmitting that information to the

TOME server. The BAN has multiple ways it interfaces with the system, including data streaming through a message broker, making requests through HTTP web services provided by the TOME server, and interfacing with sensors through various communication methods, such as Bluetooth Low Energy (BLE). The BAN is designed to serve as a gateway between the TOME server and sensors. This prevents the server from having to manage sensor connections directly, making the system more scalable and manageable. Additionally, because the BAN can remotely connect to the server (via WiFi or cellular network), users do not need to stay within a certain physical range of the server, thus carrying the potential to make them more mobile. Another benefit of the BAN transmitter is that it runs as a background service. This ensures the application is always running and maintaining an active connection with the server. It also limits user interaction with the system, which allows users to focus on their tasking without distraction.

The TOME server performs a variety of functions, including centralized real-time data processing, management of user states, management of experimental test conditions, and persistent data storage by utilizing cloud computing resources. In short, the TOME server manages all data in the system, executes algorithms, and hosts web services that interact with the system. Also running on the TOME server is a web server that provides a front-end user interface for the entire system. These applications include several different pages for viewing data, entering forms, and performing administrative functions. This also provides a convenient mechanism for exporting all collected data into a comma separated value format that can be ingested by virtually any commercial statistical software package, including both unprocessed sensor data and post-processed algorithm derived measures. Because the displays are web-based, they are compatible with a wide variety of devices, specifically any device that can run a web browsing application.

The TOME backend server includes a processing module that executes algorithms to generate alerts and derived features within the system, including inferred user states such as cognitive workload. The TOME project includes an API for algorithm development. This API includes several interfaces and abstract classes to help developers create new algorithms that can easily be used by the system. It also comes with utility methods to help evaluate algorithms in bulk. The system currently supports two main types of algorithms: 1) Data Algorithms: Algorithms that receive data messages to generate new features within the system, often referred to as "derived" features; and 2) Alert Algorithms: Algorithms that receive data messages to generate alerts within the system. Both types of algorithms fundamentally work the same way; they mostly differ with respect to the type of information they generate.

2 Experiment

10 participants (9 male and 1 female) from a University in the Western United States participated in this the study. They ranged in age from 23 to 42 years ($M = 26.4$ yrs, $SD = 8.6$) and gave informed consent under the guidelines and restrictions of the university's institutional review board. As the participants were new to the X-Plane simulator, they were given a self-paced period before undergoing the experimental conditions in which they were allowed to practice to practice until they felt comfortable with their ability to

pilot in the simulation before being allowed to move forward through the experimental apparatus.

2.1 Sensor Set-Up and Flight Simulator Testbed Evaluation

As shown in Fig. 2 (left), each participant wore all TOME sensors, which consisted of a Polar compression t-shirt embedded with electrodes around its chest cavity which connects to a Polar H10 heart rate monitor placed on the back collar of the Polar t-shirt. Respiration rate was collected using a Zephyr Bio-harness placed at the bottom of their sternum. GSR was recorded using a Polar M4500 smartwatch worn by the participants on their right wrist. Lastly a 2 channel fNIRS device, the PLUX Explorer, was placed on the participants forehead at the mid-point of their respective FpZ locations using the measured using the 10–20 system.

Fig. 2. Sensor configuration (left) and testbed environment with desk mounted Tobii Eye Tracker (right).

The experiment setup consisted of a computer equipped with an X-Plane 11 Flight Simulator installation. In order to mimic a typical in-flight cockpit of an FA-18F, the simulator was configured to simulate the same aircraft to be maneuvered with a Thrust-Master Joystick and Rudders as shown in Fig. 2.

Using pilot testing and prior literature [30–35] as guidance, we created two scenarios of high and low difficulty levels within the X-Plane environment. As depicted in Table 1, in the low difficulty level, participants were instructed to fly the plane while maintaining an altitude of $\pm$5000 ft from 10,000 ft while the weather conditions were clear and sunny, with no wind. In the high difficulty level condition, the participants had to keep their altitude within the more restrictive $\pm$500 ft from 10,000 ft, while operating under extreme weather conditions that included high wind and levels of rain.

A custom X-Plane plugin was created to enable the researchers to create experimental designs for participants, that would allow presentation of the difficulty levels (Table 1) in an order and for a duration specified by the researchers prior to the experimental protocol beginning. The plug-in also allowed researchers to build pauses between conditions

Table 1. Weather parameters and altitude constraints for low and high difficulty levels.

Parameter	Low difficulty	High difficulty
Altitude (in ft.)	10,000 ± 5,000	1,000 ± 500
Wind Speed (in kts)	55	430
Sky Conditions	Clear	Cloudy
Rain Speed (in mph)	16	110

where participants could rest. Through the plug-in, a set of instructions pertaining to the altitude instructions to follow at any given time were overlaid on the center-left side of the simulator screen during all conditions (see Fig. 3).

Fig. 3. X-Plane 11 GUI with a white arrow pointing to custom participant instructions regarding altitude and rest.

Protocol: After providing informed consent, the participants were trained on the usage of the X-Plane simulator using the X-Plane flight school. They were also shown the instructions for the plugin, which would ask them to maintain specific altitude ranges at various times. Once the participants were comfortable taking off and flying in the X-Plane environment, they were equipped with the TOME sensor suite. After take-off, each participant conducted a series of 6 'tasks'. A task represented each condition (low and high) consisting of flying for 60 s while maintaining the target altitude range amongst the weather conditions created by the plug-in. After each 60-second-long task, the X-Plane simulator would pause the screen mid-flight and participants would rest for 45 s to allow their brain's metabolic activity to return to baseline. The protocol included a randomized block design, with each block containing one low and one high difficulty level condition presented in a random order. The block design had six blocks total.

3 Data Analysis and Results

3.1 Manipulation Check

As a manipulation check, we first verified that the workload manipulations were being mentally perceived by subjects with different difficulty, we used joystick deflections from the neutral position as behavioral measure [31]. Using the first order derivative of joystick deflections ΔJ from the neutral axis we were able to perform workload manipulation check on our participant's data. Figure 4 shows the average of deflections for the first High and Low difficulty task encountered by the participants, respectively. As per the figure, the ΔJ for high difficulty tasks are significantly farther from the neutral axis *(y = 0)* than the lower difficulty setting.

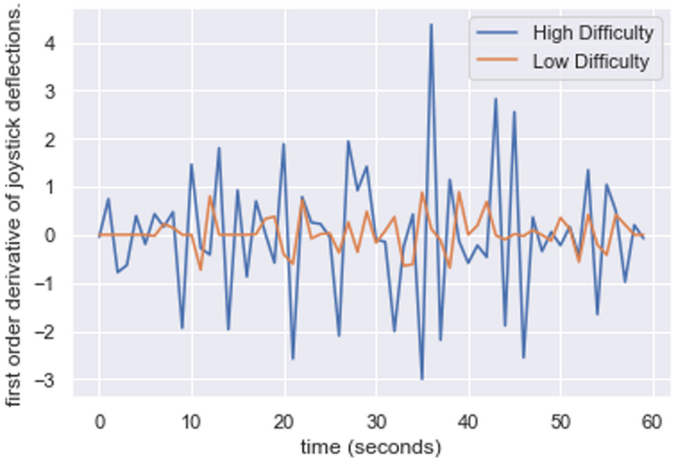

Fig. 4. Average Joystick deflections across 10 participants for the two difficulty levels.

3.2 Data Pre-processing

Since the sampling rate varies for each of the five sensors, data acquired from each sensor goes through different pre-processing routines, with outputs being placed in a raw datafile that can be exported from the TOME server. To accommodate supervised classification with feature vectors built from these five sensors, we generate all features only once every five seconds (i.e., if there are six readings from the Tobii eyetracker over a 5 s window, those will be averaged together and output as one average value for summarizing the Tobii features in that 5 s window).

Functional Near Infrared Spectroscopy: Raw data acquired from the fNIRS device was first converted into absorption coefficient I_R (μA) using the transfer function outlined below:

$$I_R = \frac{c * ADC}{2^n}$$

Here, c is a proportionality constant depending on bit precision (n) of raw data values configured within the OpenSignals Software (the default recording software for the PLUX fNIRS device). We used values 0.15 and 16 bits respectively [36]. Due to the intricate optical properties and high frequency nature of fNIRS sensors, the optical density values acquired are sensitive to the displacement of optodes from their locations resulting from head movements, micro movements resulting from cardiac pulses (Mayer Waves) or respiratory activities of the subjects [37]. We reduced motion and other physiological artifacts by applying a band-pass filter on the resulting data with values between 0.01 and 0.5 Hz. Finally, values are estimated from the two wavelengths using the Modified Beer Lambert Law to convert the data into relative changes in oxy- and deoxy- hemoglobin. This resulted in four timeseries data streams sampled at 10 Hz, consisting of Δoxy- and Δdeoxy- hemoglobin at two measurement locations. Then for every 5 s of fNIRS data we generated the following eight features (shown in Table 2): average and slope (2) x for both Δoxy- and Δdeoxy- hemoglobin (2) x two channels locations (2). These 8 features are shown in Table 2 for the fNIRS.

Heart Rate Monitor (Polar HR10): To process the raw ECG data, we band-pass filtered the raw data between 0.5 and 35 Hz followed by detrending to remove the baseline shift in the data. After extracting the QRS complex (the main spike seen in ECG data) from the filtered signal, emergence of an R peak in the data indicated a heartbeat. We then extracted 2 features from the processed data in 5 s windows: time difference between two consequent R peaks (Heartrate) and the frequency of the peaks (RR Interval).

Eye Tracking (Tobii 4C): For extracting features from the Tobii eye tracker we focused on the estimated size (in cm) of the pupil diameter, as output by the Tobii. We simply output the average diameter of the Left and Right pupil over each five second window of time (Left pupil diameter, Right pupil diameter).

Respiration Rate (Zephyr Bioharness): For the Bioharness Respiration monitor we calculated mean and standard deviations over a 5 s window, resulting in features for mean respiration rate mean and standard deviation of respiration rate.

Galvanic Skin Response and Pulse Rate (Polar M4500 Smartwatch): For the Polar smartwatch we acquired pulse rate and galvanic skin response data. We used these data streams to generate average pulse rate and average GSR activity over the course of a 5 s window.

3.3 Accounting for Missing Data and Outliers, Normalization of Resulting Data

The TOME system relies on the five devices detailed above to function 'properly' in order to collect the data to the TOME server. Issues in collecting data from a device can occur when 1) The specific device loses Bluetooth connectivity with TOME or 2) the

Table 2. List of the 16 features out every five seconds from the five sensors included in the TOME System.

Sensor	PLUX
Features	1. Average Δoxy channel 1 2. Average Δdeoxy channel 1 3. Slope Δoxy channel 1 4. Slope Δdeoxy channel 1 5. Average Δoxy channel 2 6. Average Δdeoxy channel 2 7. Slope Δoxy channel 2 8. Slope Δdeoxy channel 2
Sensor	Polar HR10
Features	9. Heartrate 10. RR interval
Sensor	Tobii
Features	11. Left pupil diameter 12. Right pupil diameter
Sensor	Bioharness
Features	13. Mean respiration rate 14. Stdev respiration rate
Sensor	Polar M4500 Smartwatch
Features	15. Pulse rate 16. GSR Activity

sensors from a given device lose contact with the participant. *Unfortunately, the PLUX and Polar Smartwatch sensors were prone to losing network connectivity, resulting in data loss during experimental sessions, as shown in* Table 3. *See 'Limitations' section for further discussion.*

Whenever a sensor did not collect data for a participant, the missing values were filled in with 'NaN' (Not a Number) values using the NumPy python package. These values were replaced so that the data from the 'working sensors' could still be used for classification while ignoring feature values resulting in the missing sensor data. This is further detailed in the sections below. The resulting data were also examined and outliers eliminated from each feature separately. For each participant and for each feature all values that were greater than 3 standard deviations away from the mean feature value were again replaced with 'NaN' values. This replacement was chosen as z-score and correlation matrix operations in many Python packages can ignore 'NaN' cells in their calculations without skewing the results. After outliers were handled the data was then normalized using z-score normalization for each of the 16 feature columns.

Table 3. Summary of device and network connectivity performance.

Participant	Description
P1	No PLUX fNIRS Readings
P2	**All Devices collecting properly**
P3	**All Devices collecting properly**
P4	No PLUX fNIRS Readings
P5	No Tobii and No Polar Smartwatch Readings
P6	No Polar Smartwatch Readings
P7	No PLUX fNIRS Readings
P8	**All Devices collecting properly**
P9	**All Devices collecting properly**
P10	No PLUX fNIRS Readings

3.4 Feature Exploration and Selection

Selection of an optimal feature set is critical to creating an efficient classification pipeline. To determine an optimal feature set, we use the oft used feature filtering method whereby Pearson correlation coefficients are generated to down-select the features that have the highest correlation with the class value [38]. We first wanted to see the relationship between the 16 features. To do this a Pearson correlation matrix was generated between every feature and every other feature across all participants and all difficulty levels. This is shown in Fig. 5.

There are some intuitive examples of correlated features in the preliminary correlations. For example, left and right pupil diameter positively correlate, and fNIRS oxy and deoxy hemoglobin have a negative correlation (which is in line with the nature of the blood oxygen level dependent signal). Furthermore, pulse rate and RR interval are correlated, which indicates that both the Polar watch and compression shirts were collecting correlated heartrate data. Aside from the obvious correlations between similar physiological sensor streams, it is promising to note that there is not a great deal of redundancy in the collected features. Strong correlations throughout would indicate redundant features, and thus give reason to down select sensors (i.e., one can increase practicality and sensor footprint by removing a sensor without losing valuable information). Of note is that it does seem that the Polar Shirt and Polar watch do have redundant heartrate information.

Next, we wanted to look at the correlation between each feature and the difficulty level (i.e., low/high). Since the feature values are continuous and the class target difficulty level is binary, we generated Kendall correlations rather than Pearson correlations, as Kendall correlation is better suited for identifying correlations between continuous variables and nominal class values. The correlation matrix shown in Fig. 6 was then generated.

As shown in Fig. 6, of note is that Left and Right pupil diameter have a strong correlation with difficulty level. Also, the standard deviation of the respiration rate also has a strong correlation. Thus, in the next section, we explore the use of linear SVMs

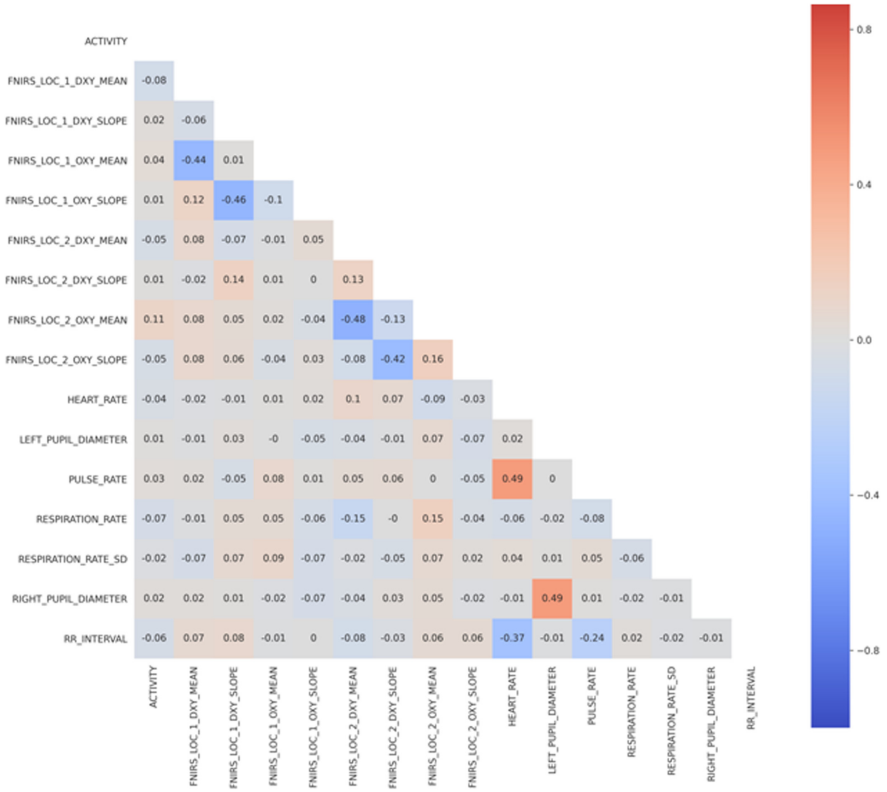

Fig. 5. Pearson Correlations between each feature.

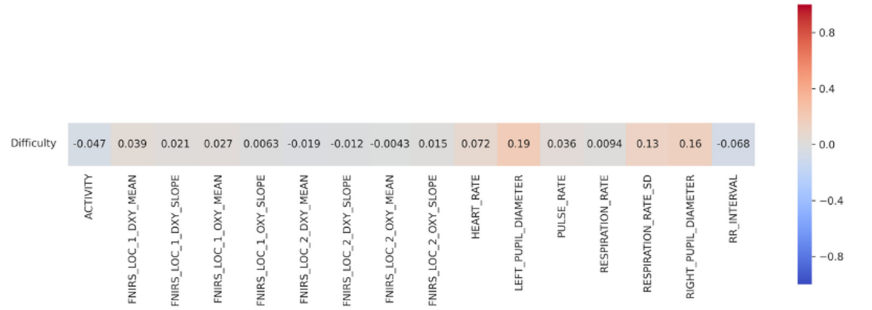

Fig. 6. Kendall correlations between each feature shown on the x axis and task difficulty shown on the y axis.

for classification using all 16 features versus classifier creation using the features that have the highest correlation with the target class value, as indicated in Fig. 6.

3.5 Classification

Based on the insights gained from the correlation matrices above, we aimed to classify our data into low vs high difficulty levels. Our goal was to explore at classification results when all features were included and when just the top features from the Kendall correlations in Fig. 6 were considered.

The z-score and correlation procedures above are robust to missing data, but most classification algorithms are not able to handle missing values. To ensure an unbiased classification, the 'NaN' generated in the previous section were replaced with values of random noise drawn from Standard Normal distribution from each of the 16 columns of features. We then opted to use a linear support vector machine to classify our data into high and low difficulty. Figure 7 shows the accuracy (left) and F1-scores (right) achieved by using all 16 features and a leave-one-participant-out cross validation scheme. In this type of cross-validation the model is trained on 9 participants and then tested on data from the unseen 10th participant. This process of train/test is repeated for each participant, and results are averaged across those 10 cross-validation runs.

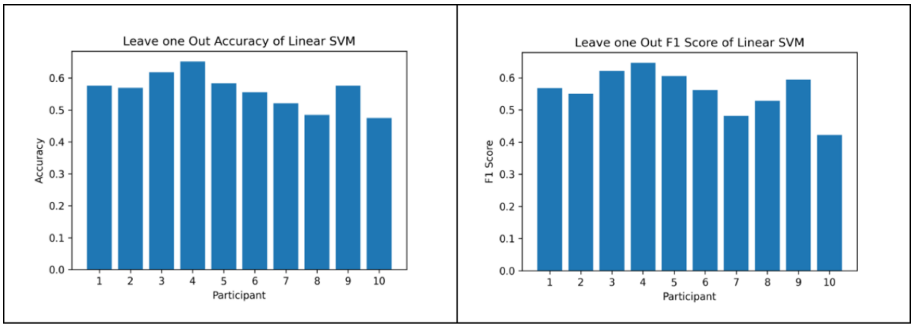

Fig. 7. Left: the figure depicts accuracy results of leave-one-out cross validation classification on the data. This dataset contained many missing data points which were filled in with noise sampled from a standard normal distribution. A linear support vector classifier from sklearn was used. The mean accuracy achieved was 56.1%. Right: depicting F1 scores. The mean F1 score was 55.8%.

The results shown in Fig. 7 are in line with our expectations. We expect lowest accuracy when we use all 16 features, especially since we know from Table 3 that many of the values fed into the model will reflect the random noise used to insert feature values where 'NaN' values are present, for the participants when a sensor did not correctly log data to the TOME server.

To see if our SVM models performed better without missing data, we ran the same SVM techniques detailed above, but for only the participants who did not have a sensor stream missing during data collection (as shown in Table 3, p2, p3, p8, p9 had all 16 features collected properly) Results are in Fig. 8, and we do note as expected that accuracy and F1 values do indeed increase when we focus on participants without missing sensor data.

The results in Fig. 8 are quite promising, as we believe that we can improve accuracy by using both performance and difficulty level as our ground truth label, as done recently

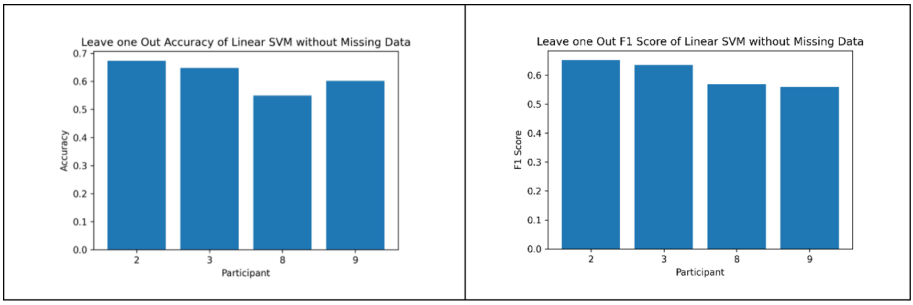

Fig. 8. Accuracy Results of leave-one-out cross validation classification on dataset containing only training examples without any missing data points. A linear support vector classifier from the sklearn python module was used. The mean accuracy was 61.8%.

by Mckendrick et al. [9]. Anecdotally we noted that participant 8 had a good deal of trouble mastering the flight simulator, as reflected in that participant's performance data.

Finally, we wanted to explore the impact that feature selection had on our model results, by only using the 'best features' from the correlation matrix in Fig. 6. As shown in Fig. 9, we select features from best to worst according to Kendall correlation results in Fig. 6. For example, when trained on only one feature, we use LEFT_PUPIL_DIAMETER, the feature with the highest Kendall correlation. When training on two features, we use both LEFT_PUPIL_DIAMETER and RIGHT_PUPIL_DIAMETER, the two features with the highest Kendall correlations. The y axis represents mean accuracy across all participants assessed in leave-one-out cross validation.

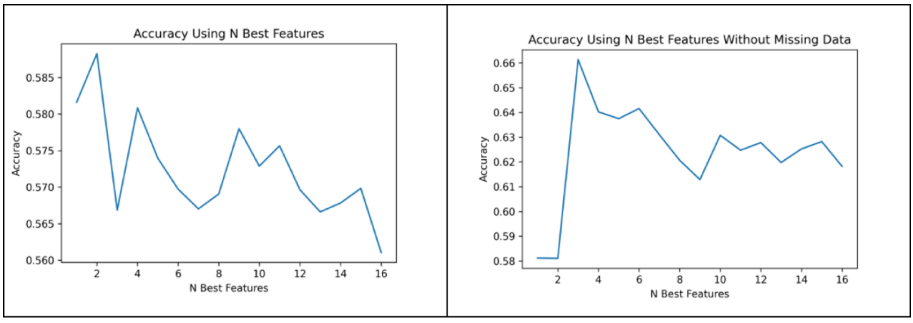

Fig. 9. left mean accuracy results for n best features. The x axis depicts the number of features used in leave-one-out cross validation classification. Right: We repeat the process used in the left, but exclude training examples with any missing features.

As noted on the right side of Fig. 9, again, we repeat the process used in the left, but exclude training examples without missing features. Here we notice an increase in mean classification with relatively few features indicating that including uncorrelated and potentially noisy features has a negative effect on classification accuracy. As shown on the right of Fig. 9, the highest accuracy achieved across all participants (using leave

one out cross validation, and all 10 participants) was just above 66% accuracy when we used only the top 3 features of Left and Right Pupil Diameter and Respiration rate. This is notable because all 10 participants had successful data collections where the Tobii and Bioharness sensors collected data properly.

4 Discussion, Limitations and Recommendations

We noted above that several of the wearable sensors would sporadically lose contact with the human subject and/or with their network connectivity to the TOME server at various times during the data collections. This would result in NaN values in their raw data. As we would expect, and as shown in the classification results above, classifications were indeed stronger when all sensors were actively collecting quality data during experiments. Below we identify these connectivity and other issues that were detrimental to our results and we talk about steps to address these issues:

- Issue 1: The PLUX and Polar Smartwatches were the main devices that would lose full connectivity during experiments.

Recommend: Finding ways to ensure more reliable connectivity for sensors. For example, the our research team has worked to bypass the PLUX OpenSignals software, and we have swapped the Shimmer GSR sensors into the TOME system as a more reliable alternative to the Polar watches. Further, we are exploring ways that data can be logged directly to a local storage device, and later uploaded to the TOME server during standard data collections in case networking issues continue.

- Issue 2: Flying flight simulators are difficult. Student participants would train on the simulator beforehand, but they still struggled to fly with a high level of comfort.

Recommend: Futures studies looking at workload need significantly more training time or to only recruit people familiar w/flight simulators. Actual pilots would be good subjects.

As detailed in Table 1, the two difficulty levels asked participants to fly within a range of 10,000 ft with a range of ±500 for low difficulty and ±5000 for high difficulty. As it has been well studied that condition level alone does not account for a proper 'mental workload label' [9], we should only include labeled data where the recorded altitude shows that participants maintained the proper range as required by the task. By merging the difficulty level with the performance, we can be more confident that our 'ground truth' labels do indeed represent the difficulty experienced at that time by a given participant.

Also noteworthy is that TOME allows us to better understand which features lead to the most information gain. The results shown in Fig. 7 are in line with our expectations. We expect lowest accuracy when we use all 16 features, especially since we know from Table 3 that many of the values fed into the model will reflect the random noise used to insert feature values where NaNs are present, for the participants when a sensor did not correctly log data to the TOME server. That said, it would be interesting to explore

different techniques for filling in data when a TOME sensor breaks down. Perhaps augmenting data from other participants (who are not currently in the test set) and adding some random noise to the resulting data could be a viable option. Also, The Tobii eye trackers right and left pupil diameter were the strongest features. This makes sense as flying in simulators is a visually intensive task and data quality was likely higher when participants were focused on the X-Plane monitors, giving a reliable eye tracking signal. It might also be possible to change the amount of time over which a feature is selected. In the above, 5 s windows of time are examined, but it is possible that varying this parameter (window size = 1 s, 10 s, 20 s) could affect classifier accuracy. Further effort should be made to explore these and other options for collating and analyzing multiple sensor streams.

Acknowledgements. We would like to acknowledge the US Navy for supporting this research via contract number N68335-18-C0133.

References

1. Borghini, G., Astolfi, L., Vecchiato, G., Mattia, D., Babiloni, F.: Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. Neurosci. Biobehav. Rev. **44**, 58–75 (2014)
2. Yuksel, B.F., et al.: Learn piano with BACh: an adaptive learning interface that adjusts task difficulty based on brain state. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, California, USA, pp. 5372–5384. ACM (2016)
3. Jimenez-Molina, A., Retamal, C., Lira, H.: Using psychophysiological sensors to assess mental workload during web browsing. Sensors **18**(2), 458 (2018)
4. Kwon, J., Shin, S., Im, C.: Toward a compact hybrid brain-computer interface (BCI): performance evaluation of multi-class hybrid EEG-fNIRS BCIs with limited number of channels. PLoS ONE **15**(3), e0230491 (2020)
5. Brouwer, A.-M., Zander, T.O., van Erp, J.B.F., Korteling, J.E., Bronkhorst, A.W.: Using neurophysiological signals that reflect cognitive or affective state: six recommendations to avoid common pitfalls. Front. Neurosci. **9**(136) (2015)
6. Lemm, S., Blankertz, B., Dickhaus, T., Müller, K.R.: Introduction to machine learning for brain imaging. Neuroimage **56**(2), 387–399 (2011)
7. Combrisson, E., Jerbi, K.: Exceeding chance level by chance: the caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. J. Neurosci. Methods **250**, 126–136 (2015)
8. Young, M.S., Brookhuis, K.A., Wickens, C.D., Hancock, P.A.: State of science: mental workload in ergonomics. Ergonomics **58**, 1–17 (2015)
9. McKendrick, R., Feest, B., Harwood, A., Falcone, B.: Theories and methods for labeling cognitive workload: classification and transfer learning. Front. Hum. Neurosci. **13**(295) (2019)
10. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. In: Hancock, P., Meshkati, N. (eds.) Human Mental Workload, Amsterdam, pp. 139–183 (1988)
11. Roscoe, A., Ellis, G.: A Subjective Rating Scale for Assessing Pilot Workload in Flight: A Decade of Practical Use. The Royal Aerospace Establishment (1990)
12. Hirshfield, L.M., et al.: This is your brain on interfaces: enhancing usability testing with functional near infrared spectroscopy. In: SIGCHI. ACM (2011)

13. Navon, D.: Resources—a theoretical soup stone? Psychol. Rev. **91**, 216–234 (1984)
14. Wickens, C.: Multiple resources and mental workload. Hum. Factors **50**(3), 449–455 (2008)
15. Berka, C., Levendowski, D.: EEG correlates of task engagement and mental workload in vigilance, learning and memory tasks. Aviat. Space Environ. Med. **78**(5), B231–B244 (2007)
16. Gevins, A., Smith, M.: Neurophysiological measures of cognitive workload during human-computer interaction. Theor. Issues Ergon. Sci. **4**, 113–131 (2003)
17. Izzetoglu, K., Bunce, S., Izzetoglu, M., Onaral, B., Pourrezaei, K.: fNIR spectroscopy as a measure of cognitive task load. In: Proceedings of the IEEE EMBS (2003)
18. John, M.S., Kobus, D., Morrison, J., Schmorrow, D.: Overview of the DARPA augmented cognition technical integration experiment. Int. J. Hum.-Comput. Interact. **17**(2), 131–149 (2004)
19. Hankins, T.C., Wilson, G.F.: A comparison of heart rate, eye activity, EEG and subjective measures of pilot mental workload during flight. Aviat. Space Environ. Med. **69**(4), 360–367 (1998)
20. Putze, F., et al.: Hybrid fNIRS-EEG based classification of auditory and visual perception processes. Front. Neurosci. **8**, 373 (2014)
21. Lohani, M., Payne, B.R., Strayer, D.L.: A review of psychophysiological measures to assess cognitive states in real-world driving. Front. Hum. Neurosci. **13**, 57 (2019)
22. Ahn, S., Nguyen, T., Jang, H., Kim, J.G., Jun, S.C.: Exploring neuro-physiological correlates of drivers' mental fatigue caused by sleep deprivation using simultaneous EEG, ECG, and fNIRS data. Front. Hum. Neurosci. **10**, 219 (2016)
23. Miller, E.E., Boyle, L.N., Jenness, J.W., Lee, J.D.: Voice control tasks on cognitive workload and driving performance: implications of modality, difficulty, and duration. Transp. Res. Rec. **2672**, 84–93 (2018)
24. Schier, M.A.: Changes in EEG alpha power during simulated driving: a demonstration. Int. J. Psychophysiol. **37**(2), 155–162 (2000)
25. Loukopoulos, L., Barshi, I.: Concurrent task demands in the cockpit: challenges and vulnerabilities in routine flight operations (2003)
26. Lancaster, J.A., Casali, J.G.: Investigating pilot performance using mixed-modality simulated data link. Hum. Factors **50**(2), 183–193 (2008)
27. Caldwell, J., Lewis, J.: The feasibility of collecting in-flight EEG data from helicopter pilots. Aviat. Space Environ. Med. **66**, 883–889 (1995)
28. Callan, D.E., Durantin, G., Terzibas, C.: Classification of single-trial auditory events using dry-wireless EEG during real and motion simulated flight. Front. Syst. **9**, 11 (2015)
29. Gevins, A., DuRousseau, D., Zhang, J., Libove, J.: Flight helmet EEG system. In: Final Tech Report AL/CF-SR-1993-0007, Sam Technology, San Fracisco, CA (1993)
30. Causse, M., Dehais, F., Pastor, J.: Executive functions and pilot characteristics predict flight simulator performance in general aviation pilots. Int. J. Aviat. Psychol. **21**(3), 217–234 (2011)
31. Boril, J., Jirgl, M., Jalovecky, R.: Use of flight simulators in analyzing pilot behavior. In: Iliadis, L., Maglogiannis, I. (eds.) AIAI 2016. IAICT, vol. 475, pp. 255–263. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-44944-9_22
32. Dorneich, M.C., Rogers, W., Whitlow, S.D., DeMers, R.: Human performance risks and benefits of adaptive systems on the flight deck. Int. J. Aviat. Psychol. **26**(1–2), 15–35 (2016)
33. Gil, G., Kaber, D., Kaufmann, K., Kim, S.: Effects of modes of cockpit automation on pilot performance and workload in a next generation flight concept of operation. Hum. Factors Ergon. Manuf. Serv. Ind. **22**(5), 395–406 (2012)
34. Lim, Y., et al.: A novel simulation environment for cognitive human factors engineering research. In: 2017 IEEE/AIAA 36th Digital Avionics Systems Conference (DASC) (2017)
35. Nocera, F.D., Camilli, M., Terenzi, M.: A random glance at the flight deck: pilots' scanning strategies and the real-time assessment of mental workload. J. Cogn. Eng. Decis. Making **1**(3), 271–285 (2007)

36. Bracken, B., et al.: Development and validation of a portable, durable, rugged functional near-infrared spectroscopy (fNIRS) device. In: Presented at the International Neuroergonomics Conference, Philadelphia, PA (2018)
37. Devaraj, A., Izzetoglu, M., Izzetoglu, K., Onaral, B.: Motion artifact removal for fNIR spectroscopy for real world application areas. In: Proceedings of the SPIE International Society for Optical Engineering, vol. 5588, pp. 224–229 (2004)
38. Koelstra, S., et al.: DEAP: a database for emotion analysis using physiological signals. IEEE Trans. Affect. Comput. (Special Issue on Naturalistic Affect Resources for System Building and Evaluation) **3**(1), 18–31 (2014)