Multi-Modal, Multi-State, Real-Time

Crew State Monitoring System

Kier Fortier
Aerospace Engineering Sciences and Bioastronautics, University of Colorado Boulder

Nikhil Garg
Electrical and Computer Engineering, The University of Texas at Austin

Elizabeth Pickering
Aerospace Engineering, Virginia Tech

Glenn Research Center
NASA Space Academy - Summer Session

Date: 09 Aug 2013

This final report has been reviewed and approved by Mentor to ensure information is accurate and does **not** contain sensitive data.

Signature_____

Mentor Name & Org Code/Date

Multi-Modal, Multi-State, Real-Time Crew State Monitoring System

Kier Fortier[1,2]
University of Colorado, Boulder, CO 80309

Nikhil Garg[1]
The University of Texas at Austin, Austin, TX 78705

and

Elizabeth Pickering[1]
Virginia Polytechnic Institute and State University, Blacksburg, VA 24061

Mentors: Dr. Beth Lewandowski[3], Angela Harrivel[3,4], Dr. Tristan Hearn[3]

This paper describes the progress toward development of a multi-modal, multi-state, real-time classification system to determine the cognitive state of an operator in safety-critical conditions. Three modalities were investigated for potential integration into such a system: electroencephalography (EEG), heart rate variability (HRV), and galvanic skin response (GSR). The entire system – including data collection, artifact rejection, data analysis, and mental state prediction – has been developed in a modular fashion. A test bed was set up with data synchronization across modalities and integration with psychological tasks. Several functional tasks were used to collect data; the most useful experiment placed a subject in a resting state, followed by a mentally engaged state, simulating inattention and attention, respectively. For each modality, various features were extracted to represent blocks of data. These features were used to first train a machine learning classifier and then predict whether a block of data originated from a resting state or a concentration state. This binary classification serves as a basis for future work toward a system that can classify a spectrum of cognitive states. Several features for each of the modalities were investigated to find the most relevant for classification, and successful features that correlate with the target mental states were found. Each of the modalities was individually used to predict subject state with an accuracy that is significantly better (with $p < 0.05$) than that of a random classifier with an input of the ratio of the amount of rest data to concentrated data. GSR and EEG features were combined into a joint classification system reaching up to 80% accuracy, more accurate than either modality alone. Real-time, multi-modal classification was also implemented for GSR and EEG, which is a major step towards achieving the overall goal of a real-time, multi-modal state monitoring system. In future work, the HRV features with the highest accuracy should be integrated into real-time, multi-modal system. Furthermore, the classifier should be tested with tasks that better simulate a pilot's operational environment. The work presented in this paper is a proof-of-concept for a real-time, multi-modal, multi-state classification system and provides a framework for such a system.

Keywords - galvanic skin response, electroencephalography, heart rate variability, cognitive state monitoring, machine learning

I. Introduction

In an increasingly computerized society, control of dangerous machinery is often automated, with a human operator checking instruments and only occasionally intervening. With auto- pilot in the commercial aviation domain, for example, a pilot does not have to worry about routine control until he or she needs to take over, perhaps suddenly, and fly the aircraft. In such situations, a pilot can become too relaxed, doze off, or zone out. On the other end of the spectrum, a novice pilot can be too anxious to safely operate the aircraft. A crew state monitoring system can detect such states and alert operators and others of the potential danger. This paper describes work done toward the development of such a system. A testbed was created to integrate multiple sensors, representative features were calculated and analyzed, and a classification system was developed to predict subject states.

[1]NASA Space Academy 2013, NASA Glenn Research Center, Cleveland OH.
[2]AIAA Student Member.
[3]NASA Glenn Research Center, Code REB, Cleveland, OH.
[4]University of Michigan, Ann Arbor

Three physiological modalities were considered: galvanic skin response (GSR), electroencephalography (EEG), and heart rate variability (HRV). EEG is a measure of voltage changes in the brain due to the flow of ions in neurons. It has historically been used for clinical applications such as epilepsy diagnoses but is increasingly being used in brain-computer interfaces (BCIs) and attention studies. GSR is an autonomic measure of skin conductivity, is typically evoked by stress, fear, or surprise, and is often used in polygraph and anxiety tests.[?] Heart rate variability is a measure of variation in interbeat intervals. Historically, HRV has been used as a strong predictor of mortality following acute myocardial infarction, but recent research has shown that it correlates with mental engagement and concentration as a physiological expression of autonomic nervous system (ANS) activity.[1,2]

II. Experimental Setup

A. Sensors

An Emotiv Epoc headset (Emotiv, San Francisco, CA) was used to collect EEG data. The Epoc has fourteen EEG electrodes and two gyros (for the X and Y directions) and is used for both research purposes and brain computer interfaces. The headset preprocesses the data before sending the data through USB: it downsamples from a sampling rate of 2048 Hz to 128 Hz , bandpass filters from 0.16 Hz to 85 Hz, and notches at 50 Hz and 60 Hz (frequencies affected by electrical noise in Europe and the United States, respectively).

For the measurement of GSR, a NeuLog GSR sensor (NeuLog, www.neulog.com) was used with a NeuLog USB hub for interface with the computer. This sensor places electrodes on the index and middle fingers of a subject's hand; it sends a very small electrical current between the nodes in order to calculate the conductivity of the skin. Skin conductance constantly changes and therefore the NeuLog GSR sensor samples at 24 Hz with a 16 bit resolution.

To detect heart rate and the electrocardiogram (ECG), a Zephyr Bioharness (Zephyr Technologies Corp., Annapous, MD) was used. A chest strap is attached around the torso just below the sternum. Sensors on the strap detect information such as heart rate, ECG, respiration depth and rate, skin temperature, and incline of the upper body. The data is then processed by the Bioharness software to provide an output of RR (interbeat) intervals, reported at a frequency of 18 Hz. For reference, the three individual sensors can be seen in Fig. 1 .

Figure 1. The EEG, GSR, and HRV sensors used in the experimental setup.

B. Functional Tasks

Various experimental tasks were set up in the lab and integrated with the sensors in order to engage a subject in varying levels of workload. These experiments were chosen to evoke different physiological and cognitive states.

The experimental procedures used in this report were approved by the NASA Langley Research Center Internal Review Board.

1. Google Earth Flight Simulator

The first functional task was the online Google Earth Flight Simulator (GEFS). Using this free software, a subject operates various aircraft over a range of locations on Earth through a joystick. The experiment was broken up into four sections: 1) five minutes of rest on the ground before takeoff; 2) one minute to takeoff and reach a steady, level flight; 3) three minutes of steady, level flight; and 4) two minutes of maneuvers such as a coordinated turn, a barrel roll, and a ground pass. Sections 1 and 3 were meant to be periods that required low mental engagement and could be defined as inattentive, and sections 2 and 4 were meant to be periods that required a high level of mental engagement and could be defined as attentive. As a proof of concept, the flight simulator was integrated with the GSR sensor and tested on several subjects.

2. Psychology Experiment Building Language

The second functional task was the Psychology Experiment Building Language (PEBL, pebl.sourceforge.net). The PEBL battery contains various psychological tasks used for numerous applications. Two experiments from the battery were considered: the Mackworth Clock Test and the Change Detection Task.[3] The Mackworth clock test consists of a small red circle moving around a larger circle, occasionally skipping the adjacent placement on its trajectory. This test is a sustained attention and vigilance task. The Change Detection task consists of many circles blinking on a screen and, for a given trial, a single circle is either changing size, color, or location. The subject must actively seek out and report the change. Accuracy, time, and type of change are among the metrics reported upon completion of the test.

3. Facilitated Rest and Engagement

The final functional task investigated incorporated the PEBL Change Detection Task sandwiched by 5 minute blocks of rest. To induce a state of focus (attention), a subject was engaged in a PEBL Change Detection task. To induce a state of rest (inattention), the subject was instructed to rest and think of nothing in particular. Data such as accuracy and time duration of the PEBL task were not considered; rather, the data taken during this experiment for each of the modalities were analyzed and input into a machine learning algorithm for binary classification of inattentive versus attentive state.

III. Methods

A. Overview

Figure 2 shows the general processing steps for the crew state monitoring system. As data is collected, it is cleaned of artifacts and noise. The signal to noise ratio (SNR) of the data should be improved as much as possible.

Next, in the most important yet difficult stage, features are extracted from the data. The goal is to calculate features that can represent each block of the data. These features should correlate with the various subject states the system is trying to detect. The data is first split into blocks. Blocks are delimited by time length, number of samples, or the serial data (identifying each PEBL trial) captured alongside the sensor data in the integrated testbed. Both the length of each block and the delimitation method are determined by the features and modalities used.

Finally, in the machine classification stage, the extracted features are used to first train a machine learning classifier (a Support Vector Machine is used) and then classify new data.

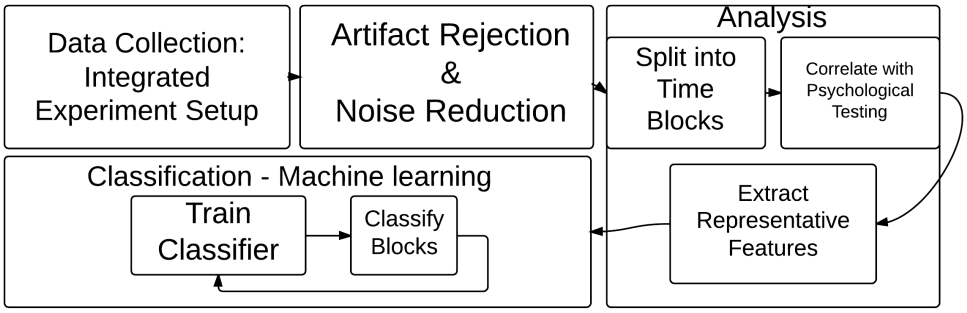

Figure 2. General processing flowchart

B. EEG

1. Background

Scalp electroencephalography, since it was first used in the 19th century, has become a common tool in the clinic. As EEG methods and tools have matured, other applications have taken advantage of its relatively cheap and noninvasive nature. In recent years, it has been used in attention studies (such as for ADHD research) and in brain- computer interfaces.[4] The Emotiv Epoc itself was originally designed for BCIs but is now also used as a research platform. However, noninvasiveness comes with the price of a low signal to noise ratio (SNR), because of which the raw data must be preprocessed and cleaned of artifacts.

Due to its long history, numerous analysis techniques have been developed to analyze EEG data. The most common EEG processing technique in attention studies literature is frequency analysis. Four main frequency ranges are considered:

delta (0 - 4 Hz), which increases during slow-wave sleep; theta (4 - 8 Hz), which increases during drowsiness; alpha (8 - 13 Hz), which increases during relaxation; and beta (13 - 30 Hz), which increases when the subject is alert. Research indicates that the power at these frequencies directly correlates with alertness level.[4] Alan Pope uses the ratio of the beta channel to the theta and alpha channels as an engagement index.[5] This ratio is often referred to as the Pope Index in this paper. However, frequency analysis requires a sensitive sensor, a low-noise environment, and a high sampling rate (from the Nyquist theorem, at least double the highest relevant frequency).

A more noise resistant technique is to find Event Related Potentials (ERPs).[6] ERPs are responses (often spikes in certain electrodes in a time series) due to specific events, such as movement, specific motor control thoughts, and blinks. They are most often used in BCIs. Due to high temporal and spatial localization and high amplitudes, event responses tend to be resistant to noise. However, though useful for seizure detections or BCI, event responses do not provide a direct measure of general brain activity and alertness level.

Less common than frequency analysis and ERPs are statistical measures such as the variance or kurtosis (measure of peakedness of data) of the data. These measures are simple to calculate. However, there is limited research available on the relation of such measures to alertness level, and no correlation may exist between these measures and engagement level.

Another potential processing method and feature source is independent component analysis (ICA). ICA is a technique in which the various source components of a signal are recovered. It can be used alongside ERPs to detect specific signal sources. In theory, signals from parts of the brain indicating the type of activity desired can be identified. However, performing ICA is computationally complex, in both code complexity and runtime, and noisy data severely complicates identifying activity strains.

2. *Open Source Tools*

Various open source tools for EEG analysis were first considered. EEGLAB, a system built on top of Matlab (but can run independently) with a sophisticated artifact removal scheme, was first used. It can reject many types of artifacts. For example, Fig. 3 shows an automatically detected blink. However, EEGLAB was not used for several reasons. Most importantly, EEGLAB does not work in real-time with the Emotiv. Data must be entered into EEGLAB after collection, and the actual removal process requires some user interaction. Second, EEGLAB rejects too much data. Even in relatively clean datasets, it rejected up to 90% of epochs in several EEG trials. Finally, using frequency analysis for artifact rejection requires the Matlab Signal Processing toolbox, which is not free. Nevertheless, EEGLAB remains a potentially useful tool for offline artifact rejection. Other available tools, such as Biosig, Pyeeg, BrainStorm, and edfbrowser, have similar shortcomings. One promising open source tool, OpenViBE, supports the Epoc Emotiv, reads and processes data in real-time, and supports Python and Matlab scripting and exporting. However, its integration was not pursued and is a suggested possible avenue for future work.

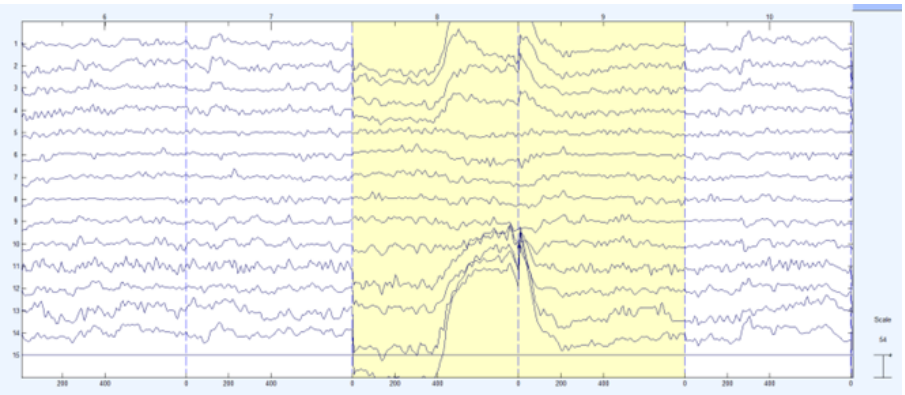

Figure 3. Automatic blink detection in EEGLAB

3. *Initial Analysis*

Initial work focused on learning how to work the Emotiv Epoc and analyzing various methods for processing. Aditya Kalluri, an intern on the project in Summer 2012, implemented EEG data reading capability, PEBL integration, and preliminary frequency analysis.[7] Figure 4 is a plot of the Pope Index over time with preliminary data. The figure illustrates the Epoc's high sensitivity to noise, especially movement noise. Outside of a two minute block in the center, the data is unusable due to movement and other noise.

Furthermore, a trade study was conducted concerning the use of wavelet transforms instead of the Fast Fourier Trans-

form. The Fourier transform is susceptible to a limitation imposed by a concept similar to the Heisenberg Uncertainty principle: time resolution directly trades off with frequency resolution. The shorter the block size for the transform, the better the time resolution but the poorer the frequency resolution, especially in the low frequencies. The longer the time blocks, the more accurate the fourier transform but the worse the time resolution. For signals with nonconstant transforms, such as EEG signals, this shortcoming could be important. Wavelets avoid the problem by using a filter bank that enables good temporal resolution for low frequency components and good frequency resolution for high frequency components. However, the fast fourier transform was still used due to the following reasons:

1. A wavelet transform does not decompose the signal into sinusoids of various frequencies. Most of the literature in EEG attention processing uses the frequency domain to analyze data, and taking advantage of existing literature was deemed a high priority.

2. Using wavelet transforms would increase both implementation complexity and runtime.

3. Sources indicated that EEG is not susceptible to the problem described above.[8] A Fourier transform is still applicable for signals in which the high frequency components vary faster than the low frequency components.

Nevertheless, wavelets transforms can be used as a feature source for classification in future work.

The initial analysis was primarily used to learn how to process the data and to make decisions for component implementations for the real- time classification system. From this analysis, an implementation plan was developed for each component of the real-time feature analysis and classification of EEG data.

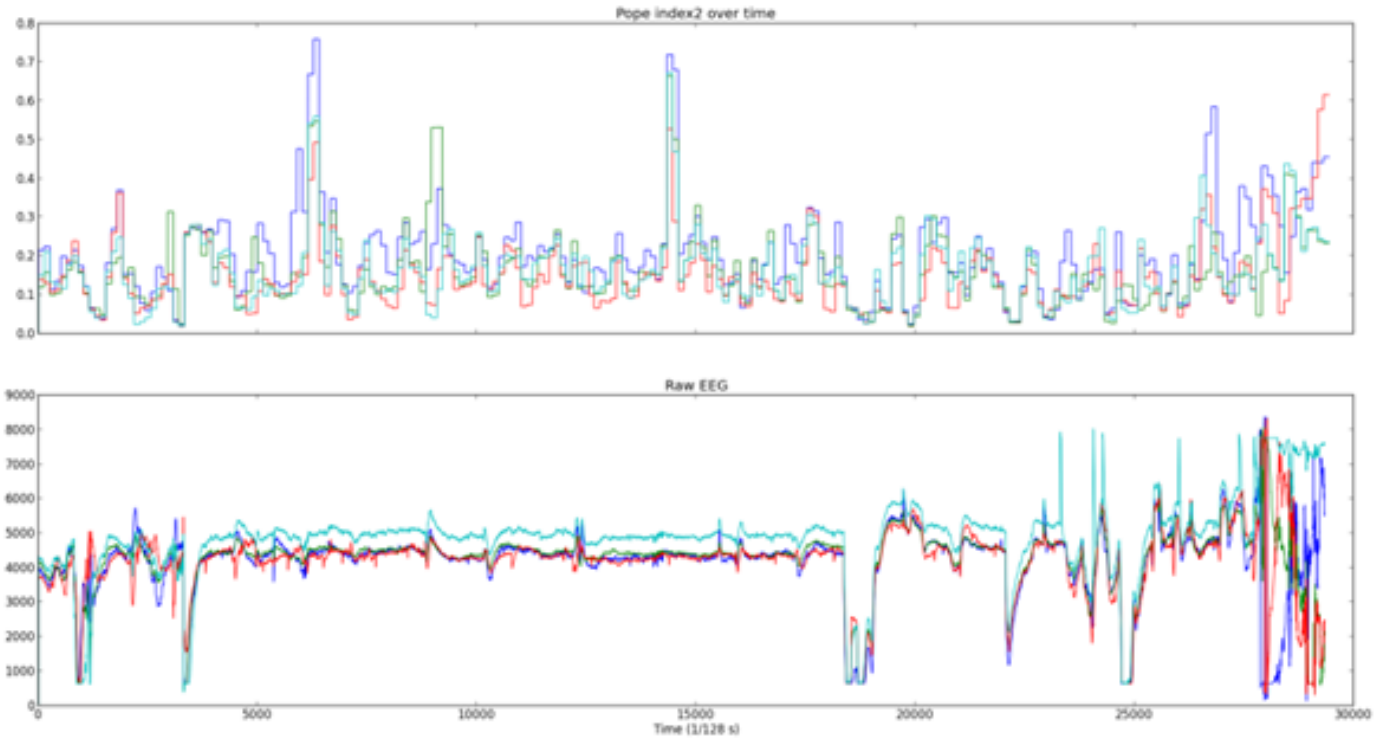

Figure 4. Initial frequency analysis

4. Artifact Detection and Removal

Before analysing EEG data, it must be aggressively cleaned of various artifacts and noise such as eye blinks, head movement, and jaw clenches. One of the greatest challenges in EEG processing is the detection and removal of eye blinks. Though blinks can be used as a feature in state detection, they are noise in respect to detecting brain activity. Eyeblinks are unpreventable (without serious implications in subject comfort), occur regularly and often, occupy a frequency range (0-4 Hz) that also contains other useful data, and drown out other components in amplitude. In clinical applications, blinks are typically removed manually, a time consuming process. For real-time applications, manual artifact removal is not an option.

After using EEGLAB for artifact detection and removal was rejected, an extensible artifact detection and removal system was developed and applied to blinks. Figure 5 is an overview of the eyeblink detection and removal process.

For each EEG channel, as the data is read, it is split into two second epochs. Each epoch is then correlated with a known eyeblink shape. To prevent boundary conditions and to detect eyeblinks across epochs, the previous epoch is included in each correlation calculation. From this correlation, peaks are chosen as the most likely locations for a blink. From these peaks, a static threshold is used to eliminate false peaks and blinks. The remaining peaks in each channel are combined, and one second windows centered at the peaks are created. These windows can cross up to two epochs, as in the correlation. Finally, each window is filtered (high pass filter from 4 - 60 Hz) to remove the low frequency blink component.

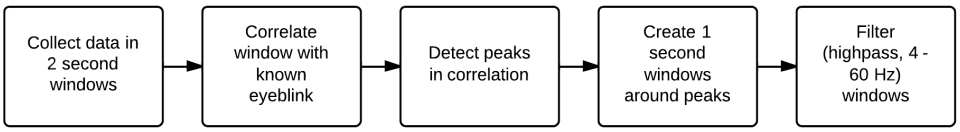

Figure 5. Eyeblink Detection and Removal Flowchart.

An identical system can be used for any known artifact shape - one simply has to replace the known eyeblink data with stereotypical data for the target artifact. A frequency based system for detection of artifacts such as jaw clenches, which have a distinctive high frequency component, has also been identified. An FFT is taken at each two second window, and the power at given frequency bins is calculated. These power values are fed into a Support Vector Machine which can be used to identify jaw clenches based on the distinctive frequency signature.

There are also alternatives to the filtering artifact removal scheme. These include rejecting the entire window around each blink (ignoring it in feature extraction and classification) and using independent component analysis to remove the blink source so that other source activities are not affected.

5. Feature Analysis and Extraction

Table 1 describes several EEG features used for data analysis.

Table 1. EEG Features

Feature	Description
Pope Index	A ratio of the power at various frequency bands. Two versions of this index are used: $\frac{\beta}{\alpha+\theta}$ and $\frac{\beta}{\theta}$. The ratios are calculated for a single channel. Time blocks of between two and ten seconds are used when this feature is used alone for classification.
Blink Rate	A function of the blink rate in each time block (ten seconds when used alone for classification). More specifically, the peaks in the correlation of the known blink and the data, as described in Section 4., are added. The sum is then normalized. This feature incorporates the four front channels of EEG data: F3, AF3, F4, and AF4.
Variance	Variance of a given channel of EEG data.
Kurtosis	Kurtosis of a given channel of EEG data. Kurtosis is the fourth standardized moment of data and is a measure of the peakedness of the data.

Of these features, the Pope Index and Blink Rate were most often used. Only the blink rate uses event responses, while the other features are long term responses to general activity. These features are used for the machine classification described in Section E.

C. HRV

1. Background

Heart Rate Variability, while a useful tool for describing ANS activity, lacks a standard method of data collection and calculation. The raw data being monitored are the RR intervals, the time between R peaks of the pulses detected by any standard ECG. There are many proposed ways of finding the ”variability” in these intervals, some of which are more or less susceptible to differences in measurement time duration. Since all of the experiments listed above have test blocks of

short, varied length (less than or equal to 5 minutes, 2 - 4 blocks), many typical methods of calculating HRV are rendered invalid. Standard deviation of intervals, for example, is a widely accepted method for calculating HRV, but depends highly on the duration of time over which the sample is taken, usually 24 hours, and should not be used to compare samples of different lengths. Other time-domain methods that could be used are the root mean square of successive differences in intervals (RMSSD) or the number of successive interval differences greater than 50 ms (NN50).[9]

Another approach is to look at the frequency-domain components of HRV. Power in the low frequency (LF) band of 0.04-0.15 Hz, high frequency (HF) band of 0.15-0.4 Hz, and the ratio between the two are often used. This is useful not only because it can be applied to test durations of varying lengths over short amounts of time, but it also indicates the type of activity going on in the brain.[9] High frequency is an indicator of parasympathetic nervous system activity (PSNS) while low frequency is less clear but is thought to be linked to both PSNS as well as sympathetic nervous system (SNS) activity. The PSNS is the part of the ANS responsible for "rest and digest" unconscious activity; this activity is depressed during concentration while SNS, responsible for the "fight or flight" response, is stimulated during high-stress or highly engaged situations. HRV, especially in the high frequencies, tends to decrease during periods of focus and so the LF/HF ratio is a good measure for classifying rest and concentration.[10]

2. *Artifact Removal*

Heart rate and HRV are affected by physical activity, but since subjects in the tests above and in anticipated experiments are kept sedentary, this was not taken into account.[10] The Zephyr Bioharness, however, is susceptible to movements as the chest strap can shift and beats can be missed or falsely detected. Besides excluding sections of especially noisy data during which the subject was adjusting the chest strap or resituating him/herself, little was done to remove artifacts from the data. The Bioharness software has its own proprietary algorithms for artifact removal and processing of the ECG. It excludes artifacts such as ectopic or incorrectly detected beats before the RR interval data is exported. However, since most software associated with heart rate monitors like the Zephyr Bioharness is proprietary, a lower-cost way of extracting and processing data in real-time would be to develop a separate code for the purposes of this project to process the raw ECG data and remove artifacts. Since the Bioharness cannot output RR data in real- time, the scope of this aspect of the project was limited. Free software packages, e.g. gHRV, for HRV analysis do exist but require data in a certain format and were not pursued. For future applications, investigation into such software as well as compatible sensors is warranted as a time and money saving measure.

3. *Feature Analysis and Extraction*

Initial data analysis focused primarily around frequency-domain measures of HRV, as described above, because this was determined to be most applicable to the data retrieved from the experiments. A set of data was taken by the Zephyr Bioharness during the Facilitated Rest and Engagement test and divided into approximately 4-minute long sections (the first and last 30 seconds for each section was excluded to eliminate adjustment time for HRV). Each block was then classified as either "rest" or "concentrated." The power spectral density was taken for each section to show the distribution differences between subject states. The power for the very low frequency (VLF, 0-0.04 Hz), LF, and HF bands was calculated, as was the ratio of LF/HF power. The power was normalized by summing the power in each frequency band and dividing by the sum of the power over the entire frequency range.

In addition, it was desirable to obtain features that could be applied over shorter time blocks (10-30 seconds) for machine learning classification. Since high and low frequency components of HRV need approximately 1-2 minutes, respectively, to accurately potray power distribution, other features were investigated. Table 2 describes the power ratio and machine learning features.

Table 2. HRV Features.

Feature	Description
Power Ratios	The ratio of power in the HF band to power in the LF band.
Mean Intervals	The average RR interval length.
Mean Power	The average power within a designated frequency band. For machine learning, the high frequency band is used.
Variance	The variance of RR intervals, a function of standard deviation.
Amplitude	The absolute value of the difference between the maximum and minimum RR interval lengths (a function of standard deviation).
Standard Deviation	The standard deviation of RR intervals.

D. GSR

1. *Background*

Galvanic Skin Response as a physiological measure has a rich and diverse heritage dating back to the early 1900's. It has been extensively used in polygraph tests, anxiety experiments, and more recently with sleep and seizure studies.[11] In general, no standard method of analysis has been developed for GSR, inherently due to its broad applicability. Although many terms are used to describe GSR, most research describes the analysis in terms of skin conductance level (SCL) and skin conductance response (SCR). SCL is a tonic measurement that refers to the value of skin conductance in the absence of any external stimuli; this can also be referred to as a baseline. SCR, on the other hand, is a phasic measurement that refers to the change in skin conductance following an external stimuli; these responses are often called peaks and are short term, occurring one to six seconds after the stimuli.[12] Common features that have been considered for the analysis of SCRs are rise time, recovery time, latency, slope, and amplitude. Recent studies have shown that the latency of response and the number of GSR peaks following a stimulus can be to indicate the state of arousal or attention of a subject.[13] SCRs are a more useful indicator for attentive states of a subject than SCL measurements, because they are a direct response to stimuli and because every subject will have a different baseline SCL measurement.

A systematic method for GSR analysis has been proposed that employs the use of principal component analysis (PCA).[12] In this publication, PCA is performed on the correlation matrix formed between the GSR test results from healthy subjects and subjects that were clinically diagnosed to be psychotic. Using various combinations of the eigenvalues of the correlation matrix, clusters were formed that were successful in separating the data between two different subject types. As the crew state monitoring system seeks to classify between various levels of attention, this research was particularly useful towards the development of a method for state separation.

2. *Artifact Removal*

The data from the NeuLog GSR sensor have artifacts that must be considered. Physical activity and hand movement were seen to produce changes in GSR that could be interpreted as SCRs. Some GSR sensors include an actigraphy measurement to account for this motion. For all subject tests performed for this report, however, motion of the hand was not necessary with the NeuLog sensor attached, and actigraphy readings could be regarded as negligible; for example, the flight simulator was controlled with the right hand while the left hand, to which the electrodes were attached, remained stationary. It was also observed that some subjects had a constantly increasing SCL throughout the duration of a test. For example, Fig. 6 illustrates this trend displayed by a subject during the flight simulation task. This trend could be correlated with an increase of palm temperature created by the contact of the skin with the joystick controller, similar to how many people develop sweaty palms while playing video games or holding hands. Small peaks can be seen in Fig. 6 which suggest SCRs exist but are not as visible due to the large change in overall SCL that occurs over the duration of the test.

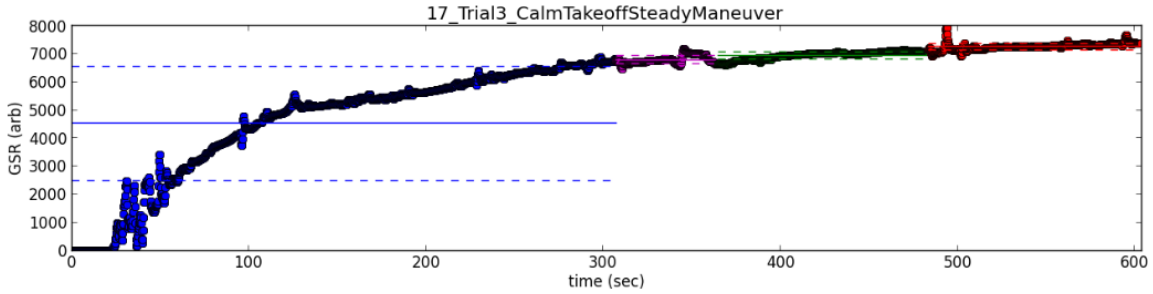

Figure 6. Increasing GSR measurements during GEFS task.

A possible solution to remove this linear trend would be to develop a temperature probe that could measure palm or finger temperature. If, for example, there was a similar increasing linear trend in temperature, that trend could be removed from the data and the relevant SCRs could be analyzed. Note that actigraphy is also evident in the first fifty seconds or so of this trial, most likely due to the subject becoming situated during setup. For an aircraft pilot, motion of the operator is unavoidable and cabin temperatures can fluctuate during flight; therefore factors such as temperature and actigraphy should be considered for a fully operational implementation of GSR in a professional crew state monitoring system. For GSR results in this report, no artifact rejection other than the minimization of subject movement during testing has been implemented. Rather than artifact removal, the GSR analysis relies on the normalization of data.

3. *Initial Analysis*

Preliminary analysis consisted of graphing the raw data and its slope at each point in time. Results for a flight simulator test can be seen in Fig. 7 , which also gives a histogram of the derivative. The histogram is centered around zero and is provided to give a general idea of whether the data is increasing or decreasing. For example, a histogram of the derivative for the results in Fig. 6 would be centered around a positive value. Figure 7 shows results similar to those obtained from the majority of test subjects. Skin conductance stayed constant or had a negative slope during the inattentive sections and had a positive slope during the attentive sections. Two subjects, however, had decreasing GSR measurements during takeoff and maneuvering phases. These subjects had the common factor of either being experienced video gamers or having flown actual aircraft. This was an important lesson in data analysis because it illustrated that habituation to stressful situations such as takeoff or maneuvers that evoked strong responses form most subjects has the potential to change the results from what may be expected; this also gave more weight to analysis methods that use features independent of the subject. Therefore, level of experience and habituation of a subject is an important factor that must be considered in developing an experiment.

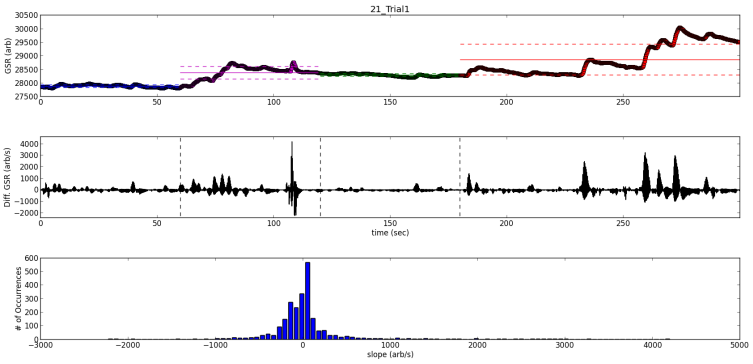

Figure 7. GSR Results for four phase GEFS scenario.

It was important to collect GSR data from multiple subjects performing different functional tasks to begin to characterize GSR shapes and trends. Subject tests were performed using the PEBL change detection task. An analysis technique was investigated using a similar PCA technique as described in the GSR background section. The results from this analysis can be seen in Fig. 8 , where the first subplot gives the raw GSR data and markers indicating the start of a new PEBL change in the task. The second subplot gives the normalized eight seconds of GSR data directly following the start of a new change detection task and the third subplot gives the average of those normalized responses at each point in time.

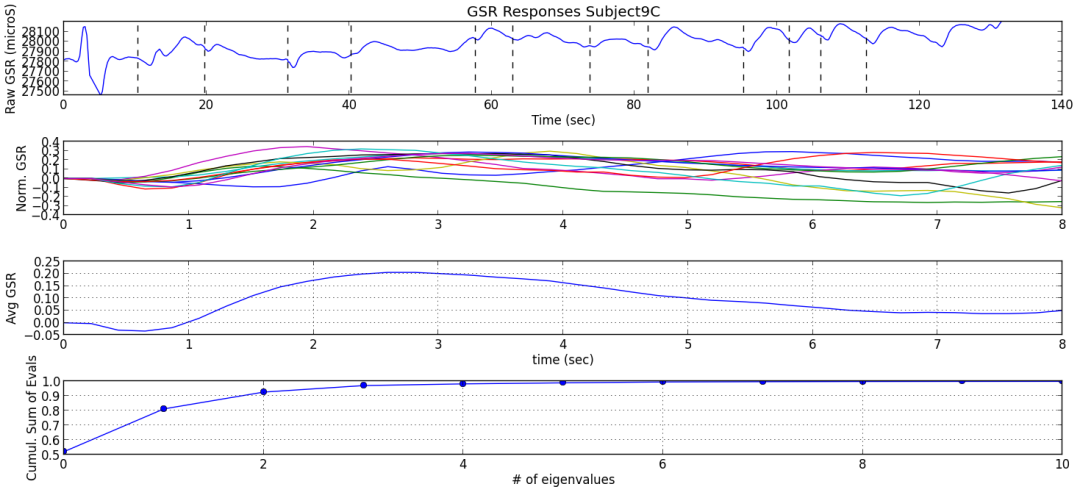

Figure 8. Sample PEBL Principal Component Analysis Results. The entire task lasted just over two minutes.

Because the responses were similar in shape and phase, the averaged value of the responses reached a normalized value of almost 0.25. For responses that are not as similar, the mean value will be closer to zero. Forming a correlation matrix for each of those events gives an idea of how similar the responses were. Each PEBL test had twelve events, leading to a 12 x 12 matrix being formed. The eigenvalues of the correlation matrix provide a way of visualizing the similarity of each response; the cumulative sum of the eigenvalues is given in the fourth subplot of Fig. 8 .[12] The sum of the first two eigenvalues for this subject is greater than 0.8, indicating strong correlation between the responses over the course of an entire PEBL test. Figure 9 gives two different clusters of eigenvalues for fourteen different subjects. The first cluster plots $\sum_{i=4}^{10}$ against $\lambda_1 + \lambda_2$ and the second cluster plots λ_4 against λ_3.

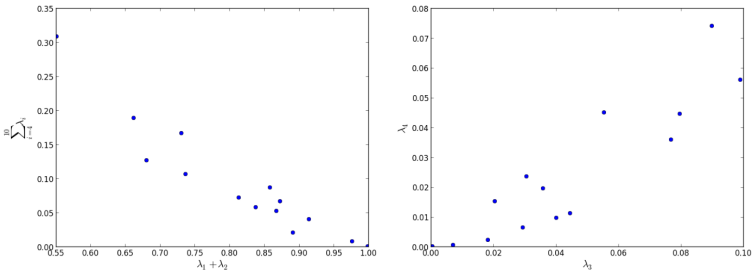

Figure 9. Correlation Matrix eigenvalue clusters for 14 different subjects from PEBL change detection task.

Although both plots appear to have roughly two clusters, further analysis yielded no correlation to attentive states. The PEBL reports generated from the program itself give metrics such as response time and accuracy but these did not correspond with the clusters seen in Fig. 9 . Since the change detection task uses four different changes of varying difficulty throughout a full test, metrics such as response time are not extremely valuable when considering the attentive state of a subject – change type is too strong of a lurking variable to overcome.

It was also observed that some subjects, while attentive, had a harder time locating changes, usually due to the subtleness of the change. Upon inspection of each set of raw data, it was evident that the correlation matrix between full responses is not extremely useful, since SCRs can have varying latency values and durations. For example, a simple increase in latency would make two similar SCRs appear different according to the correlation matrix. It was concluded that a full PEBL change detection task is not sufficient to evoke different attentive states from a subject. SCRs exist in the GSR data, but the overall shape of the data does not have any distinct feature changes. This led to the decision to implement a test consisting of five minutes of rest directly preceding and following a PEBL task, as described earlier in the Methods section.

The raw output of GSR for this task can be seen in Fig. 10 . The results for GSR from this test illustrated that GSR behaves very differently for a subject in a resting state versus a subject mentally engaged in a PEBL task. Although there were some outlying data sets that did not exactly follow this trend, in general most subjects exhibited GSR measurements similar to those seen in Fig. 10 , where the SCL is generally smooth and decreasing during the inattentive rest phase and increasing and full of SCRs during the attentive PEBL task phase. The second subplot in Fig. 10 gives a normalized average for GSR during the resting phase. These two distinct GSR shapes and trends were used to define features for the two classes, attentive and inattentive.

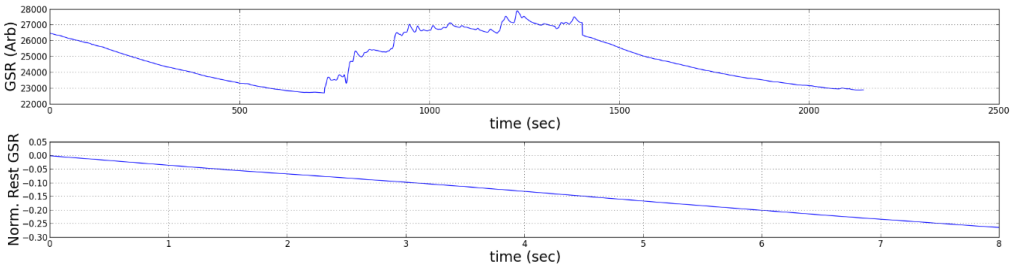

Figure 10. GSR measurement for a PEBL task preceded and followed by a resting period

4. Feature Extraction

In order to implement a crew cognitive state classification system using GSR, different existing methods were explored, leading to the development of a new technique involving feature extraction and machine learning. Table 3 summarizes

the four main features under consideration for GSR. These features were chosen based on the culmination of research on existing methods and features that best described the shape and trend of GSR data. Because SCL varies subject to subject, it was important to consider features that were independent of amplitude or the SCL. In addition to the features listed in Table 3 , features such as mean and standard deviation were calculated.

Table 3. GSR Features.

Feature	Description
Variance	Variance of the GSR data.
Max Difference	The largest difference between adjacent GSR data points.
Amplitude	The absolute value of the difference between the maximum and minimum of the GSR data.
Integral	The area under the curve formed by the GSR data, obtained by a Simpson's numerical integration scheme.

E. Machine Learning Classification

A generic machine learning code was developed to analyze features from each of the sensors. The resulting modularity allows quick testing and use of different combinations of features across modalities. The primary method used is a Support Vector Classifier from the SkLearn Python library [14]. Support Vector Classifiers (SVC) rely on the features having different values depending on the group to which they belong. Training data is first used to fit the classifier and find separations in the data. Figure 11 demonstrates such a classifier with various kernel types. The kernel refers to the type of function used to separate the data intro groups. LinearSVC and SVC with Linear Kernel are two mechanisms that use linear functions to separate groups. The RBF kernel is used for circular groups, such as when one group is completely surrounded by another. Finally, the Polynomial kernel uses a 3rd degree polynomial to separate groups. Though the examples shown only have 2 features (plotted on the X and Y axis, respectively), a SVC can support any number of features. The dots' colors represent the group to which the training data belongs. Future data is plotted similarly and classified using the separations. The background colors indicate the groups to which future data will be classified.

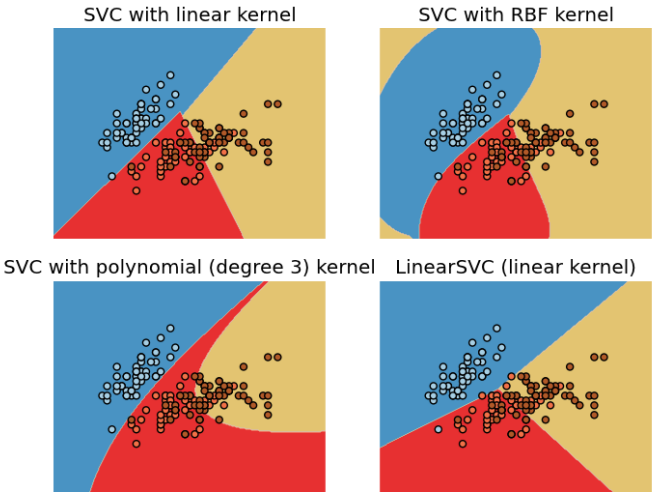

Figure 11. SVC Example with 4 different kernel types.

The Facilitated Rest and Engagement Experiment described in Section II.B.3. can be used to test and analyze the effectiveness of the classifier. Rest and engaged data from multiple subjects is used, and the data is marked according to the set from which it originates. After features are extracted from the modalities, a cross correlation mechanism is used. 60% of the data is randomly selected to act as training data. After the classifier is fit, the remaining 40% of the data is predicted, and the predicted states can be compared to their true states. After the classification stage, multiple metrics are used to evaluate the classifiers. In addition to total classifier accuracy, precision and recall for both rest and concentration were determined. Precision means, out of the things identified as X, how many were truly X, and can be calculated as

follows:

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \tag{1}$$

Recall is defined as the number of things identified as X that were truly X, and can be calculated as follows:

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \tag{2}$$

IV. Results and Analysis

A. EEG

1. *Artifact Detection*

The eye blink detection system shows promising results. It was tested as follows: with 3 subjects, two male and one female, tests were run in which a test examiner marked each subject blink with a keyboard press. These presses were saved alongside the EEG data. Qualitative analyses indicate that the system detects about 80% of blinks. A two- sample means T-test over several tests indicate that the difference between the eyeblink correlation in the one second windows around true eyeblinks and outside such windows is significant at $p < .001$.

Figure 12 is a sample output of the blink detection and rejection code. It includes the raw eeg values, true subject blinks, blink correlation and peaks, and the cleaned EEG.

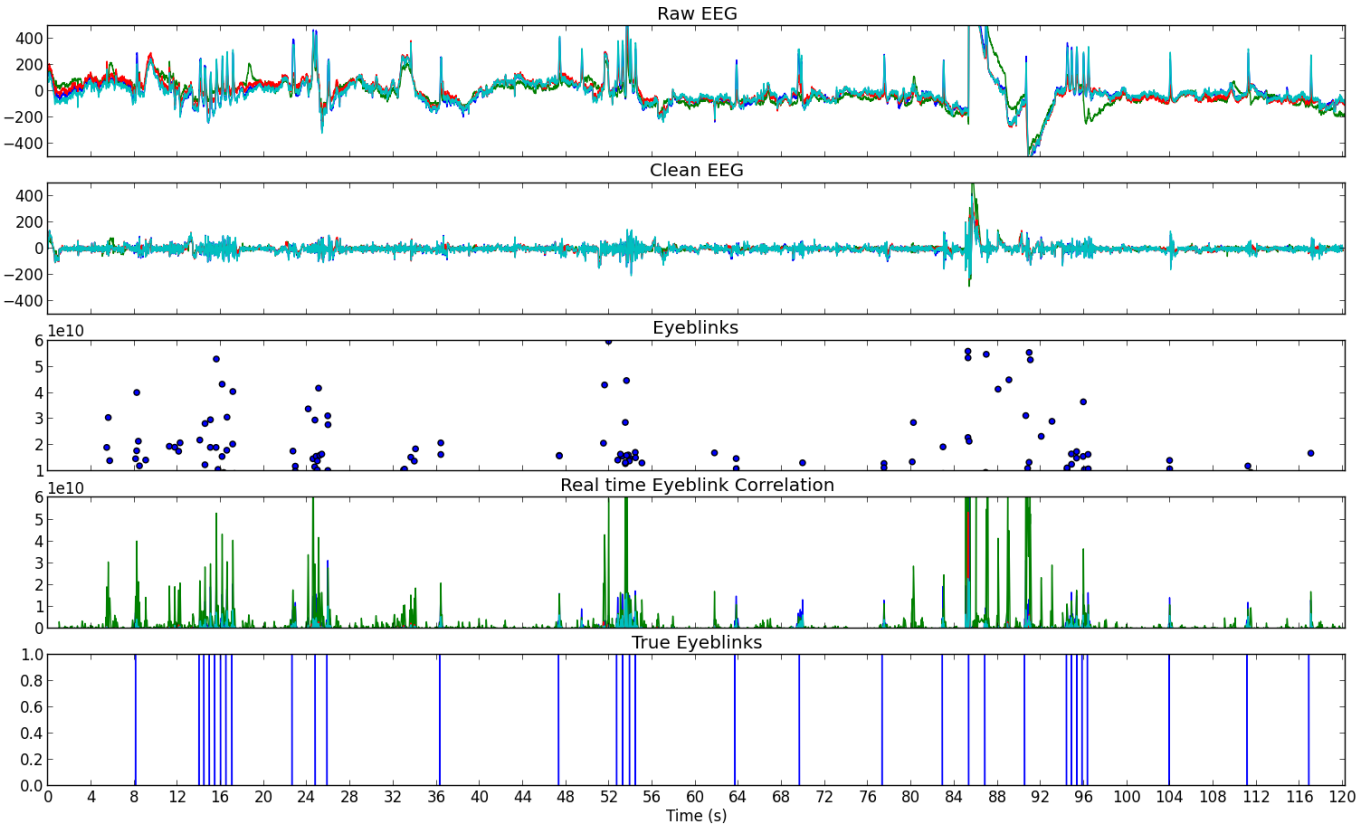

Figure 12. Blink Detection and Rejection Graphs

Furthermore, the windows around each blink were analysed for a more quantitative evaluation of blink detection. The keyboard presses and correlation peaks were used to determine, at each time step, whether the data should have been filtered and whether the data was filtered. Table 4 shows the results from this analysis and its sensitivity to the threshold

value used for peaks. These results indicate a high detection rate (of the data points that should have been detected, the percentage that was detected) but also a higher than desired false detection rate (of the data that should have been ignored, the percentage that was detected) for low threshold values. Figure 13 illustrates the True Detection and False Detection rates over increasing threshold. As the threshold increases, both the detection rate and the false detection rate fall.

Table 4. Blink Detection Results

Trial	Threshold	Data Points	True Detect	False Detect	True Miss	False Miss	Detection Rate	False Detection Rate
1	.1e10	13325	1161	8249	3708	207	84.87%	68.99 %
1	.2e10	13325	873	6486	5471	495	63.82%	54.24 %
1	.3e10	13325	707	5486	6471	661	51.68%	45.88 %
2	.1e10	15965	1532	5981	7928	524	74.51%	43.00 %
2	.2e10	15965	897	3671	10238	1159	43.63%	26.39 %
2	.3e10	15965	462	2837	11072	1594	22.47%	20.40 %
3	.1e10	16217	1385	7216	7381	235	85.49%	49.43 %
3	.2e10	16217	1125	5606	8991	495	69.44%	38.41 %
3	.3e10	16217	1006	4600	9997	614	62.10%	31.51 %
Total	.1e10	45507	4078	21446	19017	966	80.85%	53.00 %
Total	.2e10	45507	2895	15763	24700	2149	57.39%	38.96 %
Total	.3e10	45507	2175	12923	27540	2869	43.12%	31.94 %

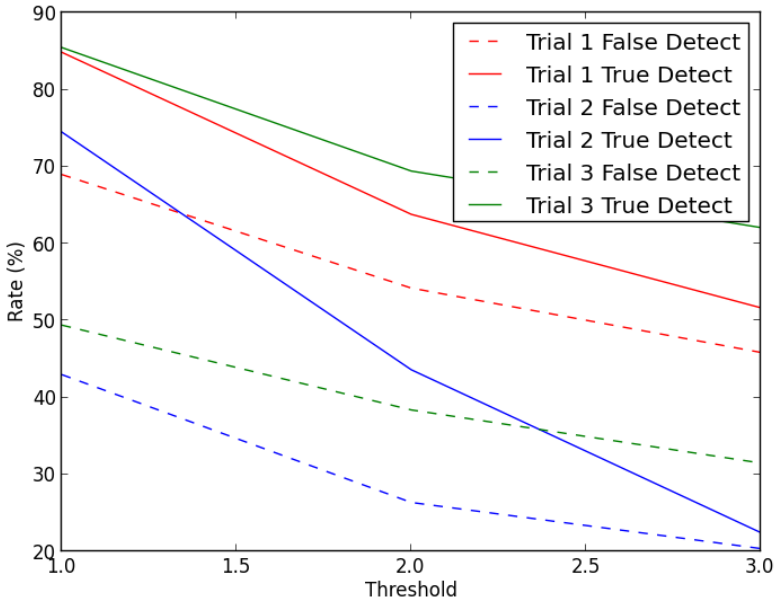

Figure 13. Detection Rates at varying thresholds

To improve this detection scheme, the threshold used to determine true artifacts from the peaks must become more sophisticated. Due to differences between subjects and experiments, as well as limitations in manual threshold determinations, a static threshold is not optimal. Regardless, the analysis indicates that the correlation technique effective in detecting features. Future work should expand this system to other artifacts.

2. *Features and Classification*

Features were used for classification both alone and in pairs. For each run, total accuracy as well as other measures were used. Eight trials of Facilitated Rest and Concentration from five subjects, two male and three female, was used for

these tests. As described in Section III.E., a random 60% of the data was used as training data, and the remaining 40% was used as samples to predict. The four kernel types mentioned above were used for classification. Table 5 contains classification results using only EEG features. It is important to note that the data was skewed with 40.9% of the data collected during the PEBL task and 59.1% of the data collected during rest. Thus, a classifier attains a success rate of 59.1% even if it classifies everything as rest due to no separation in the data. To test whether the classifier's performance over such a non-classifier is statistically significant, a hypothesis test is performed over the binomial distribution as follows:

$$P(Y \geq y|n\,,\,p) = \sum_{k=y}^{\infty} \binom{n}{k} k^p (n-k)^{1-p} \tag{3}$$

where
n = Number of time blocks in test data
p = Accuracy of a non-classifier, max(% rest data, % concentrated data)
y = classifier_accuracy_score*n

The resulting value is the probability that, if the classifier is no better than a classifier as effective as the non-classifier described above, it would yield the resulting accuracy score. The p value is thus:

$$p\_value = 1 - P(Y \geq y|n\,,\,p) \tag{4}$$

Table 5. EEG Classification Results

Feature 1	Feature 2	Best Classifier	Acc.	p_value	Rest Precision	Rest Recall	Conc. Prec.	Conc. Recall
Pope	Eyeblink	RBF	.684	.0096	.67	.93	.77	.32
Pope	None	Linear SVC	.643	.091	.70	.68	.56	.59
Pope	Variance	Linear SVC	.643	.091	.70	.68	.56	.59
Pope	Kurtosis	Linear SVC	.643	.091	.69	.73	.57	.52
Eyeblink	Kurtosis	Polynomial	.604	.368	.62	.87	.54	.22
Eyeblink	None	Polynomial	.604	.368	.62	.87	.54	.22
Variance	Kurtosis	Polynomial	.597	.432	.59	1.00	1.00	0.02
Eyeblink	Variance	All Same	.591	.497	.59	1.00	0.00	0.00
Variance	None	All Same	.591	.497	.59	1.00	0.00	0.00
Kurtosis	None	All Same	.591	.497	.59	1.00	0.00	0.00

When using the low performing features, all the test data was classified as resting because no separation could be determined from the training data.

These results suggest the potential of using EEG features in a classification system. The Pope Index performs the best, regardless of the second feature. A more sophisticated power analysis along with a more sensitive sensor would yield a powerful classifier. Furthermore, these results emphasize the benefits of using multiple features. Though the eyeblink feature alone did not perform significantly better than the baseline, combining it with the Pope index improves performance over only using the Pope index. The classifier is better than random with $p < .01$.

3. Artifact Removal

Artifact removal was tested alongside features and classification. Features were calculated using EEG from which blinks have been removed. Table 6 contains classification results using EEG features calculated from clean EEG data.

Table 6. Clean EEG Classification Results

Feature 1	Feature 2	Best Classifier	Acc.	p_value	Rest Precision	Rest Recall	Conc. Prec.	Conc. Recall
Kurtosis	None	Polynomial	.604	.368	.62	.85	.53	.25
Pope	Eyeblink	RBF	.597	.432	.61	.87	.52	.21
Pope	None	All Same	.591	.497	.59	1.00	0.00	0.00
Variance	None	All Same	.591	.497	.59	1.00	0.00	0.00

Features calculated after removing eyeblinks performed substantially worse than features calculated before artifact rejection. The high false detection rate and imperfect filtering remove too much information from the data. Future work should include finding alternatives to remove artifacts without affecting the rest of the data. These alternatives include Independent Component Analysis and individual channel filtering. Furthermore, open source tools such as OpenViBE may be used for feature rejection.

4. Discussion & Future Work

A complete EEG classification system, from data collection to feature calculations to classification, has been developed. Due to the modular nature of the code, different implementations of each component can easily be swapped into the overall system. Future work should use the developed system and simply improve various components as necessary. The results suggest that frequency power analysis has a strong potential in state monitoring systems. A classifier using the Pope Index as a feature performed significantly better than a random classifier with $p < .01$. For future work, the Pope Index feature can be improved by factoring in the Index at multiple channels. For the classifications above, only the AF3 electrode was used to calculate the Pope index. A calculation using more electrodes may prove more robust. In addition, numerous other potential features, such as wavelet analysis, ICA, and more advanced measures, can be pursued, and the machine learning classification system can also be potentially improved. Though SVMs are commonly used for such applications, other approaches such as neural networks and hidden Markov models may increase performance and should be studied.[15]

The artifact detection results display the power of a simple and flexible system to detect known responses through a correlation method, and the results support the use of the Emotiv Epoc to detect event responses such as blinks. However, from both the results and a qualitative assessment, the Emotiv Epoc is most likely neither sensitive nor noise resistant enough for long term attention research. To make further advances in feature calculations, a better sensor with more electrodes should be used.

B. HRV

1. *Features and Classification*

Initial data for four subjects, two male and two female, was taken for the Facilitated Rest and Engagement Test and the power spectral density was found and plotted. No strong patterns between subjects were apparent by visual inspection. However, distinctions between resting and concentration can be seen on the individual subject level.

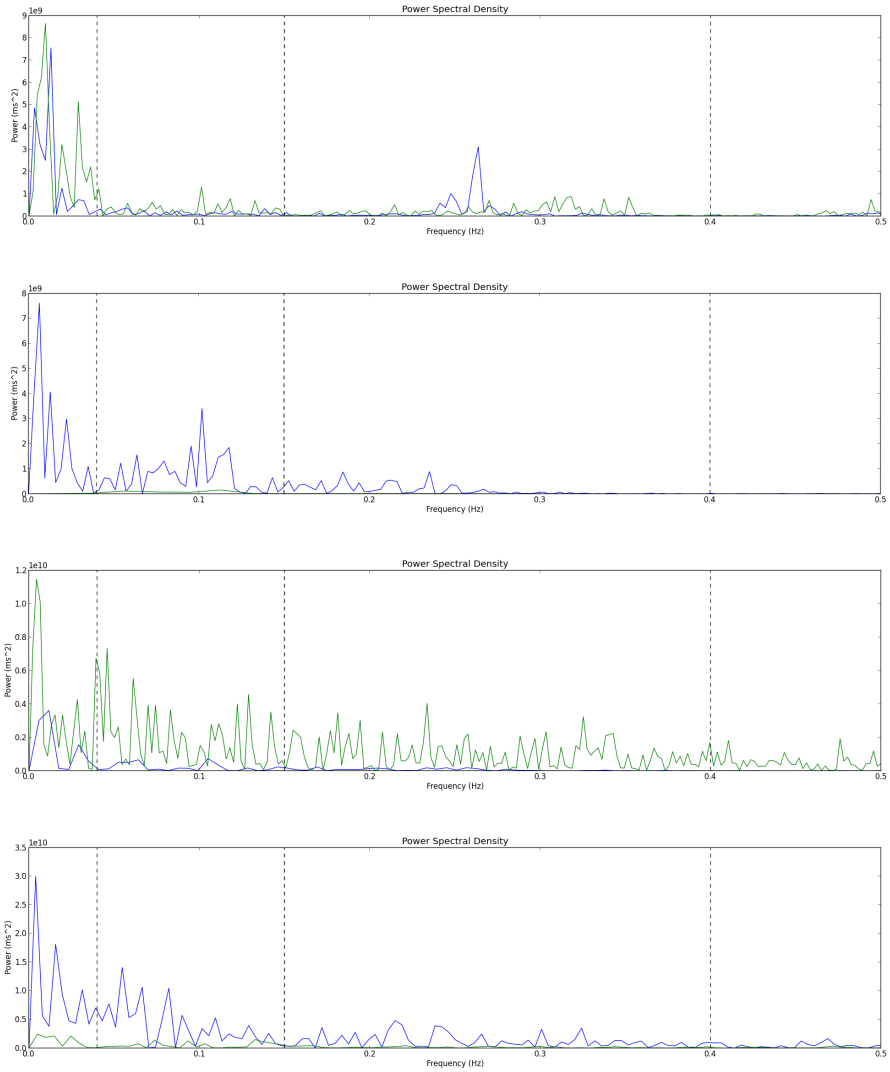

Figure 14. Power Spectral Density. From top to bottom: Subject 1, Subject 2, Subject 3, Subject 4. Blue lines represent resting data, green lines represent concentrated data, and dashed vertical lines show divisions between power bands.

Table 7. HRV Power Ratios.

Subject Number	LF/HF Resting	LF/HF Concentrated
1	2.13	2.12
2	2.39	3.80
3	3.44	1.66
4	2.93	1.67

Table 8 above shows the features used for machine learning classification as presented in EEG Features and Classification. Equation 4 was used to calculate the binomial distribution statistics in the table, where 60% of the data was used to train the classifier and 40% was used to predict state. The true percentage of resting data was 54%. Below in Fig. 15

Table 8. HRV Classification Results

Feature 1	Feature 2	Best Classifier	Acc.	p-value	Rest Prec.	Rest Recall	Conc. Prec.	Conc. Recall
Standard Dev.	Amplitude	RBF	0.707	0.013	0.61	1.00	1.00	0.45
Variance	Amplitude	RBF	0.707	0.013	0.61	1.00	1.00	0.45
Mean Intervals	Variance	Polynomial	0.683	0.029	1.00	0.32	0.63	1.00
Mean Power	Variance	Linear SVC	0.683	0.029	0.62	0.79	0.76	0.59
Amplitude	None	RBF/Poly.	0.659	0.059	0.59	0.89	0.83	0.45
Variance	None	RBF	0.659	0.059	0.58	1.00	1.00	0.36
Mean Intervals	Mean Power	RBF	0.659	0.059	0.59	0.89	0.83	0.45
Mean Intervals	Amplitude	RBF	0.659	0.059	0.59	0.89	0.83	0.45
Mean Power	None	RBF	0.634	0.11	0.57	0.89	0.82	0.41
Mean Power	Amplitude	RBF	0.634	0.11	0.56	0.95	0.89	0.36
Mean Power	Stand. Dev.	RBF	0.634	0.11	0.57	0.89	0.82	0.41
Standard Dev.	None	Polynomial	0.585	0.18	1.00	0.11	0.56	1.00
Mean Intervals	Stand. Dev.	Polynomial	0.537	0.39	0.50	0.32	0.55	0.73
Mean Intervals	None	Linear/RBF	0.463	0.84	0.46	1.00	0.00	0.00

can be seen the machine learning classifiers for the four significant feature combinations ($p < 0.05$).

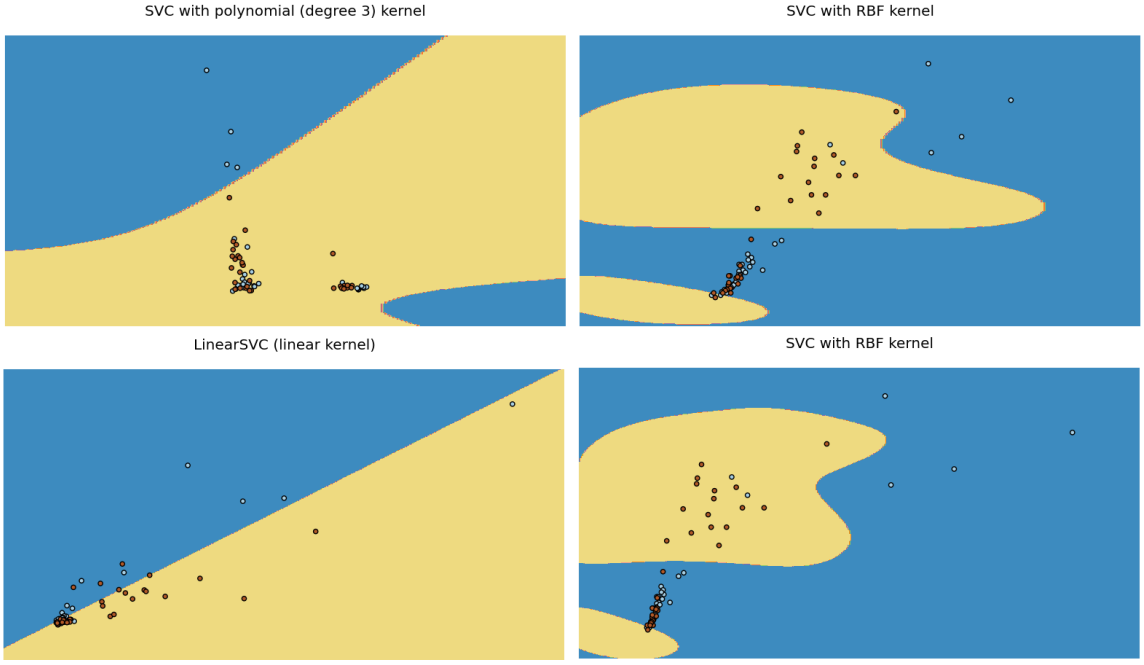

Figure 15. HRV Machine Learning Classifiers. Counterclockwise from top left: (Mean Intervals, Variance), (Mean Power, Variance), (Variance, Amplitude), (Standard Deviation, Amplitude). Red dots represent resting data and blue dots represent concentrated data; yellow background indicates resting classification and blue background indicates concentrated classification.

2. *Commentary and Conclusion*

The difference between resting and concentrating power spectral density data indicated that although there may not be a common in-task HRV response among subjects, there was a noticeable difference between attentive and inattentive states for each. Since using this method to classify attention requires much more data and more subjects than for machine learning classification, it was not pursued further. Without statistical analysis it could not be concluded whether the differences in the power ratios were significant. There was not enough data to determine whether there was a trend among power ratios, but further investigation into this may be warranted, as well as an investigation into common factors among subjects that may contribute to LF/HF increasing or decreasing with mental load, such as physical fitness.[10]

Variance or standard deviation, in combination with other features, contributed to all of the significant classifiers ($p <$

0.05). Since all the blocks of data are the same time duration, it is appropriate to use standard deviation and its derivative variance as a measure of comparison. Amplitude also appeared to be a helpful feature, nearly significant (p = 0.059) by itself, and improving performance of the predictor when paired with other features. The mean interval length performed the worst of the features (p = 0.84), making it worse than a random classifier, but mean power (taken in the HF band), while not significant by itself, did predict with a p-value of 0.11. Given that the blocks of data were 30 seconds in length,shorter than the recommended sampling time of 1 minute to detect power in the HF band, this accuracy may improve with larger time blocks. However, more data collection would be necessary to ensure an appropriate number of blocks are available for analysis.

In conclusion, more effort was put into the machine learning classification than power spectral density since it bears the potential to be easily integrated with GSR and EEG in joint classification, as will be discussed in the next section. In order to best be compatible with the time block samples for GSR and EEG classification, features that do not require long blocks to be meaningful (e.g. mean power) should be used; variance, standard deviation, and amplitude are a few such features. If longer times and more data are available, frequency analysis is appropriate and carries potential as a classifier. Though more research into the best classification system needs to be conducted, significant headway into identifying relevant features has been made and the foundation for including HRV features relatively easily in a joint classification system has been laid.

C. GSR

1. Features and Classification

After much preliminary analysis was done, the general behavior and shape of GSR for two different cognitive states were obtained. The features described in Table 3 were calculated for various subjects in attentive and inattentive states throughout the course of an entire test. The results of the machine learning method developed for this project can be seen in Table 9 . It is evident that the integral of the data is the feature that leads to the best separation of the data. All three machine learning classifications that employed the integral had a p value of $4.8\text{x}10^{-6}$, calculated with Eq. 4 , indicating that the classifier defined by the training data was much more significant than a random classifier. Note that the other three features have p values of 0.43, indicating that they are poor features to use in a machine learning classifier. Indeed, when they are combined with the integral there is no improvement to the classifier and the integral is clearly the best performing feature in terms of separation.

Table 9. GSR Classification Results

Feature 1	Feature 2	Classifier	Acc.	p-value	Rest Prec.	Rest Recall	Conc. Prec.	Conc. Recall
Integral	None	Linear	0.80	$4.8\text{x}10^{-6}$	0.97	0.68	0.68	0.97
Integral	Variance	Linear	0.80	$4.8\text{x}10^{-6}$	0.97	0.68	0.68	0.97
Integral	Max Change.	Linear	0.80	$4.8\text{x}10^{-6}$	0.97	0.68	0.68	0.97
Max Change	None	Linear	0.59	0.43	0.59	1.00	0.00	0.00
Variance.	None	Linear	0.59	0.43	0.59	0.92	0.42	0.08
Amplitude.	None	Linear	0.59	0.43	0.59	1.00	0.00	0.00

Running the machine learning code for the integral features resulted in a strong separation centered about zero; features from the attentive state had positive areas and features from the inattentive state had negative areas. A few data points do not follow this trend, however the majority of the test results behave according to this separation. The integral is a much stronger feature than the others given in Table 9 : the integral of the normalized response curve describes the trend of the data and whether the section of data being analyzed has an increasing or decreasing trend. Additionally, the other features are more prone to noise and are not as accurate over smaller sections of time such as those being considered to extract features.

2. Commentary and Conclusion

As a sensor, the NeuLog GSR is sufficient for the purposes of a crew state monitoring system. It interfaces well with Python and data can easily be extracted through the serial port, essential for a real-time implementation of state classification. Its 16 bit resolution is sufficient to capture changes in GSR that may arise from changes in a subject's cognitive state, although artifacts such as actigraphy and skin temperature may need to be considered for a more advanced system.

After much research on existing analysis techniques and technologies, a new GSR analysis method was developed

for use in state classification using machine learning. This method focused on the shape and trend of GSR data and successfully defined the integral of a data set as a feature that performed well for machine learning classification between attentive and inattentive states.

D. Multi-Modal Classification

In addition to independent feature calculations and classification, EEG and GSR were combined in a joint classification system. The data was synchronized through the integrated testbed. For each ten second block, both GSR and EEG data was retrieved and sent to their respective feature analysis functions. The returned features were then sent to the classification code. Table 10 contains the joint classification results.

Table 10. Multi-modal Classification Results

EEG Feature	GSR Feature	Best Classifier	Acc.	p_value	Rest Prec.	Rest Recall	Conc. Prec.	Conc. Recall
Pope	Area	Linear SVC	.805	$5.86\text{x}10^{-9}$	.96	.70	.69	.95
Eyeblink	Area	Linear Kernel	.799	$1.71\text{x}10^{-8}$	.97	.68	.68	.97
Pope	Variance	Linear SVC	.643	0.091	.70	.68	.56	.59
Pope	Max Change	Linear SVC	.643	0.091	.70	.69	.56	.57

The first two entries in Table 10 illustrate that the joint EEG- GSR binomial state classification system is a significant improvement over a random classification system. A combination of Pope index for EEG and area for GSR gives a p value well under the 0.01 significance threshold. Although the final two entries in Table 10 are above the 0.05 significance threshold, the combination of GSR and EEG features improves the classification accuracy and significance over the classification using the modalities individually. Figure 16 , Fig. 17 , and Fig. 18 demonstrate how using joint classification and multiple features in general improves the prediction rate. The Pope Index by itself has significant, but limited, effectiveness. When used with the GSR area under the curve, it helps separate values that the GSR alone could not handle, such as those in the top right of Fig. 18 .

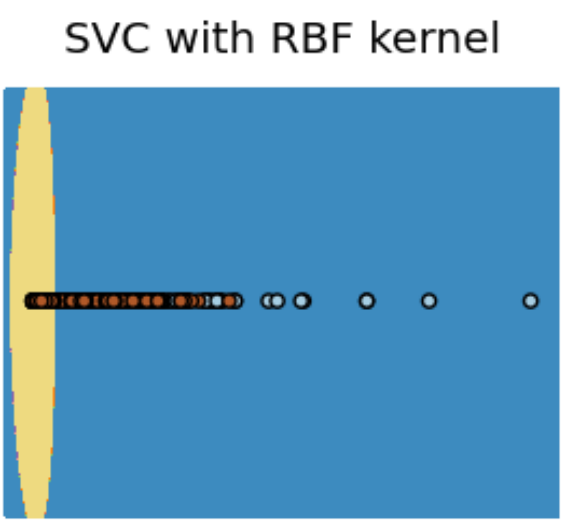

Figure 16. Classification using only the Pope Index. Red circles represent concentrated data and blue circles represent rest data.

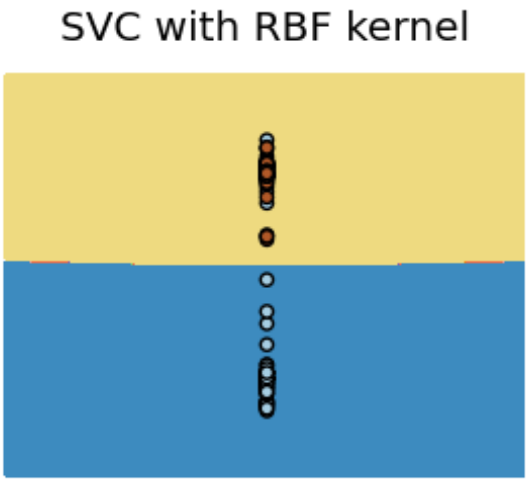

Figure 17. Classification using only the GSR area under the curve. Red circles represent concentrated data and blue circles represent rest data.

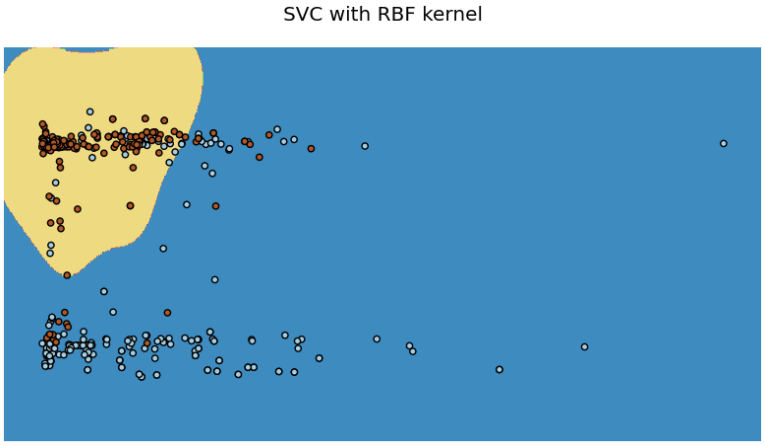

Figure 18. Classification using both the Pope Index and the GSR area under the curve. Red circles represent concentrated data and blue circles represent rest data.

The same classification system was used such that data from the same subject was used both to train the classifier and test the classifier. The accuracy score for the results from one subject increased to 87%, indicating a classifier trained for a single subject has the potential to be more accurate. The accuracy score for a second subject, however, decreased to 73%. These results illustrate that training the classifier for one subject can increase classification performance but that one must be careful to collect enough training data from the subject. Individualizing training data drastically reduces the amount of data used to train the classifier and so could reduce classifier performance. A commercial airline implementation of the crew state monitoring system could be used on the same pilot for multiple flights, justifying time spent to individualize training data.

E. Multi-Modal, Real-Time Classification

Finally, as a proof of concept for real-classification, an overall system has been developed. This system trains a classifier using previously collected data, collects new EEG and GSR data in blocks of ten seconds, calculates the features, and outputs a prediction for the subject state. The predictions are determined based on a voting function that uses the four different types of kernels. If a majority of the classifiers predict the same state, that state is chosen as the prediction. However, if the classifiers are split evenly, the Linear SVC's predicted state is chosen due to its high performance.

To test the real-time classifier, two subjects, one male and one female, switched between states of rest and concentration as instructed by a test examiner. Figure 19 shows a screenshot of the real time classification results alongside a PEBL test used to evoke concentration. Initial qualitative evaluations indicate that the system effectively predicts the subject's state. For example, when one subject was completing the PEBL task, all but two blocks over a two minute span (twelve blocks) were predicted as concentration data. However, no quantitative evaluations were made to test the effectiveness of the voting mechanism and the system as a whole.

Figure 19. Screenshot of real-time classifier in action

This system indicates that a real-time crew state monitoring system using GSR and EEG is possible. Future work should improve the voting mechanism (potentially weighing the classifiers and the features based on their individual performance), implement the ability to view signal quality (and classification confidence) alongside the prediction, and integrate this system with the real-time visualization system.

V. Project Conclusion

This report presents the work toward the development of a crew state monitoring system capable of monitoring multiple cognitive states using biofeedback from multiple modalities. A testbed was developed that integrated various functional tasks with electroencephalography, heart rate variability, and galvanic skin response sensors. Extensive research was done on existing analysis methods in order to identify meaningful features for each modality during different cognitive and physiological states. These features were extracted from subject data obtained during a functional task that evoked both attentive and inattentive states. Feature extraction from each of the modalities enabled the training and use of a machine learning binomial classifier. Classifiers using EEG Pope Index, GSR Area under the curve, and HRV amplitude and variance, respectively, performed significantly better than a random classifier, indicating that these modalities would contribute value to a crew state monitoring system. Furthermore, multi-modal classification was implemented for the GSR and EEG sensors. Of the different feature combinations, GSR area and EEG Pope index had the best accuracy score of up to 80% for multiple subjects, better than the features alone performed. This result indicates that the multi-modal characteristic of such a system yields higher performance. Finally, a real-time classification system was implemented and tested on two subjects as they were instructed to alternate between resting and concentrating. The classifications coincided very well with the actual state of the subject, demonstrating a proof of concept of a real-time, multi-modal, multi-state crew state monitoring system.

Acknowledgments

The authors would like to thank their mentors, Dr. Beth Lewandowski, Angela Harrivel, and Dr. Tristan Hearn for the enormous amount of support, guidance, and knowledge they provided. The authors would also like to thank AmandaMarie Adams from George Fox University, Newburg, OR, for her work with real-time visualization of the physiological data and setup with the test bed and Elise Eiden from University Notre Dame, Notre Dame, IN, for her work with PEBL, testing procedure, and test bed setup. Also, the entire CSM team appreciates all the volunteers that came in for various tests. Finally, the authors would like to thank the NASA Space Academy and their fellow academites for a memorable summer.

References

[1] Luciano Bernardi, Joanna Wdowczyk-Szulc, Cinzia Valenti, Stefano Castoldi, Claudio Passino, Giammario Spadacini, and Peter Sleight. Effects of controlled breathing, mental activity and mental stress with or without verbalization on heart rate variability. *Journal of the American College of Cardiology*, 35(6):1462 – 1469, 2000. ISSN 0735-1097. doi: http://dx.doi.org/10.1016/S0735-1097(00)00595-7. URL http://www.sciencedirect.com/science/article/pii/S0735109700005957.

[2] Antonio Luque-Casado, Mikel Zabala, Esther Morales, Manuel Mateo-March, and Daniel Sanabria. Cognitive performance and heart rate variability: The influence of fitness level. *PLoS ONE*, 8(2):e56935, 02 2013. doi: 10.1371/journal.pone.0056935. URL http://dx.doi.org/10.1371%2Fjournal.pone.0056935.

[3] N. H. Mackworth. The breakdown of vigilance durning prolonged visual search. *Quarterly Journal of Experimental Psychology*, 1(1):6–21, 1948. doi: 10.1080/17470214808416738. URL http://www.tandfonline.com/doi/abs/10.1080/17470214808416738.

[4] L. Eugene Arnold. Introduction: Eeg brain waves: A wave of the future or past? *Journal of Attention Disorders*, 17 (5):371–373, 2013. doi: 10.1177/1087054713485422. URL http://jad.sagepub.com/content/17/5/371.short.

[5] Alan T Pope, Edward H Bogart, and Debbie S Bartolome. Biocybernetic system evaluates indices of operator engagement in automated task. *Biological Psychology*, 40(12):187 – 195, 1995. ISSN 0301-0511. doi: http://dx.doi.org/10.1016/0301-0511(95)05116-3. URL http://www.sciencedirect.com/science/article/pii/0301051195051163.

[6] Steven J. Luck. *An Introduction to the Event-Related Potential Technique (Cognitive Neuroscience)*. A Bradford Book, 1 edition, August 2005. ISBN 0262621967. URL http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0262621967.

[7] Aditya S. Kalluri. Attentional state classification via eeg-based engagement index thresholding. *NASA Glenn Research Center Internship Report*, 2012.

[8] Robi Polikar. The wavelet tutorial - part iii: Multiresolution analysis & the continuous wavelet transform, 1996. URL http://users.rowan.edu/~polikar/WAVELETS/WTpart3.html.

[9] Heart rate variability: Standards of measurement, physiological interpretation, and clinical use. *Electrophysiology, Task Force of the European Society of Cardiology the North American Society of Pacing*, 93(5):1043–1065, 1996. doi: 10.1161/01.CIR.93.5.1043. URL http://circ.ahajournals.org/content/93/5/1043.short.

[10] Caroline Di Bernardi Luft, Emlio Takase, and David Darby. Heart rate variability and cognitive function: Effects of physical effort. *Biological Psychology*, 82(2):186 – 191, 2009. ISSN 0301-0511. doi: http://dx.doi.org/10.1016/j.biopsycho.2009.07.007. URL http://www.sciencedirect.com/science/article/pii/S0301051109001549.

[11] M.Z. Poh, T. Loddenkemper, N.C. Swenson, S. Goyal, J.R. Madsen, and R.W. Picard. Continuous monitoring of electrodermal activity during epileptic seizures using a wearable sensor. *Conf Proc IEEE Eng Med Biol Soc*, pages 4415–4418, 2010.

[12] M. P. Tarvainen, A. S. Koistinen, M. Valkonen-Korhonen, J. Partanen, and P. A. Karjalainen. Analysis of galvanic skin responses with principal components and clustering techniques. *IEEE Transactions on Biomedical Engineering*, 48(10):1071 – 079, 2001.

[13] W. F. Prokasy and D. C. Raskin. *Electrodermal Activity in Psychological Research*. Academic Press, Inc., New York, NY, 1973.

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[15] F Lotte, M Congedo, A Lcuyer, F Lamarche, and B Arnaldi. A review of classification algorithms for eeg-based braincomputer interfaces. *Journal of Neural Engineering*, 4(2):R1, 2007. URL http://stacks.iop.org/1741-2552/4/i=2/a=R01.