Classifying Stress in a Mobile Environment

Sara Dechmerowski, Brent Winslow, George Chadderdon, Tarah N. Schmidt-Daly, David Jones
Design Interactive Inc.
Oviedo, FL
sara@designinteractive.net, brent.winslow@designinteractive.net

ABSTRACT

Over half of all Veterans suffer from stress-related illnesses; of particular concern is PTSD. In addition to supporting post-deployment stress treatment, it is critical to integrate stress inoculation training pre-deployment to teach proper coping mechanisms and prevent the PTSD cycle from starting. A challenge with developing such training is the objective, real-time monitoring of stress across trainees. Current methods for stress monitoring are laboratory-based (not mobile), and episodic in nature (e.g. self report). Wearable physiological sensors provide a quantitative assessment of stress (such as heart rate variability and electrodermal activity); however, the main challenge with these technologies is the lack of robust algorithms to classify stress in a mobile environment in real time. Physiological sensors are often activated by other inputs such as temperature and physical activity, and individual differences (e.g. age, gender, health status) and daily activities (e.g., physical movements, environmental changes, caffeine intake) pose a complex problem in achieving an accurate classifier. A review of several stress monitoring algorithms published in literature has been conducted and applied to a study designed to collect the high quality data necessary for modeling and development of a classifier that accurately detects stress in a mobile environment in real-time. The study procedures, results, and development of this algorithm are outlined, including use of unobtrusive hardware and robust logic to disseminate between psychological stress and physical activity. Although the main objective of developing a mobile classifier using non-invasive sensors to classify stress with over 85% accuracy was achieved, further refinement is needed to maintain the high level of accuracy across a variety of users and environmental conditions. Future research will include further accuracy refining through reduction in environmental noise and a smart algorithm to learn individual user stress thresholds. Applications for this research within the military and others are discussed.

ABOUT THE AUTHORS

Sara Dechmerowski is a Research Associate at Design Interactive Inc. in the Human Systems Integration group. She has over 5 years experience in training systems design and evaluation, performance assessment, and next generation human systems integration.

Dr. Brent Winslow is Lead Scientist at Design Interactive, Inc., and has over 10 years of experience in studying the central nervous system at multiple scales from a bioengineering perspective. Prior to joining Design Interactive, Brent did post-doctoral work at the Allen Institute for Brain Science in Seattle WA in biosensing devices.

George Chadderdon is a Senior Systems Modeler at Design Interactive, Inc. and has previously designed and developed pattern classifiers for speech recognition, phoneme segmentation, engine vibration conditions, and communication signal types.

Tarah N. Schmidt-Daly is a Research Associate at Design Interactive, Inc. She received her MSc degree in Modeling and Simulation from the University of Central Florida in 2013.

David Jones is a senior research associate at Design Interactive, Inc. He has served as technical lead and Principal Investigator for efforts focused on designing medical training platforms that provide advanced performance metrics and adaptive real-time support systems for patients and healthcare providers.

Classifying Stress in a Mobile Environment

Sara Dechmerowski, Brent Winslow, George Chadderdon, Tarah N. Schmidt-Daly, David Jones

Design Interactive Inc.

Oviedo, FL

sara@designinteractive.net, brent.winslow@designinteractive.net

INTRODUCTION

Background

The prevalence of stress-related illness in military personnel, especially when returning from deployment, is alarming. In fact, over half of all Veterans suffer from some type of psychological distress, including, but not limited to post-traumatic stress disorder (PTSD), major depressive disorder (MDD) and suicide (NAMI, 2014). While not an overt physical injury, psychological distress is equally as debilitating. Of particular concern is PTSD, as it has been termed the 'signature wound' of the current military conflicts (RAND, 2008). Further, both diagnostic levels of PTSD and sub-threshold levels are related to poor quality of life, including anger, stress, alcoholism, depression, poor physical health, and increased suicidality (Yarvis & Schiess, 2008; Marshall, et al., 2001). In addition to treating PTSD following deployment, efforts to prevent the onset of mental illness are also critical. Research has shown that inducing stress during training and teaching coping mechanisms will better prepare Warfighters to manage the stress of combat and can prevent the PTSD cycle from starting (Follin, 1994).

There are a number of strategies and simulations for inducing stress during military training (e.g. Bartone, 2006; Carroll et al, 2012; Jones et al, 2012; Saunders, Driskell, Hall, & Salas, 1996); however, methods of inducing stress, the timing and delivery of stressors, and targeted coping strategies are all important considerations in delivering effective stress training (Kavanagh, 2005). Although moderate levels of stress have been shown to enhance performance (Russo, et al., 2012), chronic and acute stressors may degrade performance and lead to long term negative outcomes, such as those experienced by Warfighters during life threatening situations- as evidenced by the mental health crisis in the Department of Defense (DoD; Pitman, et al. 2012). Given this, it is critical that Warfighters be trained to effectively operate under stress while building resilience to chronic stressors to prevent the onset of PTSD. A challenge with developing such training is the objective monitoring of stress across trainees. This is difficult to accomplish during training because it requires subjective, post-training confirmation of the occurrence of stress.

To support the need for real-time objective stress monitoring for both stress training applications and post-deployment psychological health treatment throughout the military, wearable physiological sensors and associated mobile applications for data visualization hold enormous potential. Wearable physiological sensors can provide a quantitative assessment of stress (such as heart rate variability and electrodermal activity) (Malik et al., 1996; Sun et al., 2012), and mobile applications can support remote monitoring, and alert the wearer or instructor of real time changes in stress. Detecting and addressing chronic stress is a key measure for mobile health applications, but the main challenge in addressing this need is the lack of robust algorithms to classify stress in a mobile environment in real time. There is a growing need to support the classification of stress and other physiological and psychological states in a natural environment, but wearable physiological sensors are also activated by other inputs such as temperature and physical activity. Individual differences (e.g. age, gender, health status) and daily activities (e.g., physical movements, environmental changes, caffeine intake) pose a complex problem in achieving accurate classification of stress. The approach outlined in this paper is aimed at mitigating these challenges to obtain a reliable, accurate classifier of stress using unobtrusive sensors.

Stress Monitoring Algorithms

Existing studies developing algorithms for stress detection tend to use a wide array of features calculated from sensor data measuring various aspects of heart-beat, respiration, and skin perspiration, which are all highly responsive to increased sympathetic nervous system activity, which correlates with stress and vigilant arousal

(Everly & Lating, 2013). Particular sensor technology useful in classifying stress include ECG (Plarre et al., 2011; de Santos Sierra et al., 2011; Sun et al., 2012), skin conductance measurement (Alamudun et al., 2012; Bakker, Pechenizkiy, & Sidorova, 2011; Choi, Ahmed, & Gutierrez-Osuna, 2012; de Santos Sierra et al., 2011; Sun et al., 2012), and measurement of chest volume for respiration (Choi, Ahmed, & Gutierrez-Osuna, 2012; Plarre et al., 2011). Heart-beat features used include various measures of heart-rate and heart-rate variability (Malik et al., 1996). Skin conductance features include overall skin conductance level and phasic skin conductance responses (Choi, Ahmed, & Gutierrez-Osuna, 2012; Sun et al., 2012). Respiration features include breathing rate and respiratory sinus arrhythmia (Plarre et al., 2011). Most successful methods of classifying stress break the data stream into short periods of time, typically on the order of minutes, and attempt to classify stress vs. non-stress, and sometimes also the context of physical activity, e.g., whether the subject is sitting, standing, or walking (Sun et al., 2012). Generally, standard supervised machine learning methods—for example, decision trees, support vector machines, or Bayesian networks—are used to develop the classifiers, which means that subject data needs to be collected where the subjects are engaged in tasks known to induce stress so that "ground truth" stress or non-stress labels can be assigned to the feature vectors. Previous work has emphasized the difficulties imposed on stress classification by individual subject differences in physiological responses to stress (Alamudun et al., 2012; de Santos Sierra et al., 2011), which makes it difficult to build subject-independent stress classifiers. Another concern is the physical activity of subjects (Alamudun et al., 2012; Sun et al., 2012), which triggers many of the same physiological signals as psychological stress, leading to masking and confounds of stress detection.

Opportunity

Stress monitoring algorithms have generally been built with obtrusive physiological sensors and laboratory based settings that do not translate well to operational settings (Alamundun et al., 2012; Plarre et al., 2011), such as military training and post-deployment mental health treatment. New to market hardware options such as Empatica's E3 system or the Basis Health Tracker band are both wrist-worn devices with associated mobile applications that have the potential to take real-time stress monitoring outside of the laboratory and into everyday life. There is an opportunity to combine foundational mobile stress monitoring algorithm research methods with new to market, high quality, and unobtrusive physiological sensor suites to create an accurate, quantitative classifier for continuous and objective real-time stress assessment.

It is hypothesized that with the creation of appropriate experimental conditions and associated data collected, a valid, reliable, classifier can be developed using non-invasive sensors that predicts times of increased stress in one minute intervals with over 85% accuracy.

METHODS

Participants

Thirty-five participants ranging from 18 to 51 years (24 men, 11 women, M_{age} = 25.7, *SD* = 6.18) were recruited for the study. Participants were recruited using recruitment flyers posted online and through recruitment fairs located at local universities. Recruitment materials contained study information and requirements (i.e. age range, basic education, etc.). Informed consent was received from each participant before the study began. Experimental procedures and protocols were in accordance with the governing institutional review board. Participants were compensated for their time.

Tasks

The study consisted of several tasks that have been found to experimentally produce a stress response (excluding viewing of happy videos) based on previous literature. These tasks were used to create "ground truth" states of baseline rest (not stressed), physical stress, psychological stress, and co-occurring physical and psychological stress. A description of each task is below:

Treadmill Combination (treadmill plus mental arithmetic) Task – Participants walked on a treadmill at 3.0 mph for a total of 10 minutes. During the first five minutes of the task, the participants walked in silence. During the last five minutes of the task, the participants walked while also sequentially subtracting a single digit odd number from a four digit even number, and gave their answers aloud. The researcher alerted the participant, if a mistake was made, to begin again from the first four digit even number (Roth, Bachtler, & Fillingim, 1990).

Trier Social Stress Test (TSST) – Participants first completed a five minute speech preparatory period where the participant was left alone in a separate test room to prepare a speech. The participant then gave their speech to research panel members and also a video-camera (assumed to be on and live streaming to individuals trained in public speaking). Finally, participants completed a mental arithmetic task where participants sequentially subtracted a double digit odd number from a four digit even number and gave their answers aloud. The researchers remained neutral throughout the TSST and alerted participants when there was more time to fill or that an incorrect answer was given during the math portions (Kirschbaum, Pirke, & Hellhammer, 1993).

Grip Strength Task – Participants performed a grip strength task, continuously, for one minute. A hand dynamometer was used to induce a physical task capturing hand strength, recorded in pounds per square inch of pressure. An average maximum voluntary contraction (MVC) was obtained at the beginning of the study. During the one minute grip strength task, participants were to hold and maintain the device at a third of their MVC (Vijayalakshmi, Madanmohan, Bhavanani, Patil, & Babu, 2004).

Anger-Inducing Video Clips – Participants watched two short film clips used in emotion research to elicit anger; one clip from *Once Were Warriors* (Jenkins & Andrewes, 2012), and one clip from *Schindler's List* (Schaefer, Nils, Sanchez, & Philippot, 2010). Total time of both clips was 5 min 6 sec.

Happiness-Inducing Video Clips[1] – participants watched two short film clips used in emotion research to elicit happiness; one from *For the Birds*, a Disney short (Bartolini, 2011), and one clip from *Remember the Titans* (Bartolini, 2011). Total time of both clips was 8 min 1 sec.

Apparatus

Two sensor suites were selected for data collection: Empatica's E3 band (www.empatica.com) and the Biopac MP150 system (www.biopac.com). These two systems were selected to ensure the stress algorithm is compatible across multiple hardware solutions. Participants were fitted with both physiological sensors at the wrist and finger locations that collected a combination of pulse plethysmography (PPG) activity, electrodermal activity (EDA), temperature, and movement (3-axis accelerometer).

Procedures

Participants first completed their informed consent form and then a questionnaire packet at the reception area. Questionnaires used included the Patient-Reported Outcomes Measurement Information System (PROMIS) – Anger; the Positive and Negative Affect Schedule expanded (PANIS-X); the Depression, Anxiety, and Stress Scale (DASS); and a demographics form. Participants then moved back into the study room and were assigned to one of five groups. Each of the five groups performed tasks in a different, counter-balanced order (see Table 1).

Once group assignments were made, participants were fitted with both sets of physiological sensors (E3 band and Biopac MP150 system). Participants were then read the study overview and were also told that they would be answering the verbal Subjective Units of Distress Scale (SUDS) rating question after each task and their responses would be recorded in an experimental log. Participants verbally reported their SUDS rating for a subjective experimental check of perceived stress after each task was completed. Participants were instructed to verbally report, on a scale of 0 to 100, with 0 indicating that they were completely relaxed and 100 indicating that they were experiencing severe stress.

[1] It should be noted that the data collected from the viewing of happiness-inducing video clips was not analyzed for inclusion in the stress classifier and is beyond the scope of this paper. Subsequent analyses are intended in future research.

Next, participants completed a physiological baseline in which they rested for a period of five minutes in order to obtain baseline measurements of their physiological state. They were instructed to sit quietly and to limit movement of their arms and legs. They could rest their eyes but were told to not slouch or lie down.

Then, depending on group assignment, cortisol samples were collected on a subset of participants. Cortisol was collected on a subset of participants as an experimental check for the induction of stress after the baseline rest phase (Cortisol baseline) and then again after the administration of the TSST followed by Anger-Inducing Videos (Groups 1 and 3 had this order; Dickerson & Kemeny, 2004). Saliva samples were collected, via passive drool, into a collection tube. Total time for cortisol collection varied per participant, based on the quality of saliva in the collection tube (i.e. 1 ml of saliva with no bubbles).

Next, the experimental tasks were completed in the order outlined in Table 1, followed by rest periods, and finally the participants were debriefed and the entirety of the study was explained. Participants were compensated for their time and left with a debriefing form with additional contact information.

Experimental Design

A true random counter-balancing between task orders was not achieved due to time and space constraints However, task order was balanced through a Latin-square design for n=5.

Table 1. Experimental design: Counter-balanced order of tasks.

Task Order*	Group 1	Group 2	Group 3	Group 4	Group 5	
1	Treadmill Combo	TSST	Grip Strength	Angry Videos	Happy Videos	
2	Happy Videos	Treadmill Combo	TSST	Grip Strength	Angry Videos	
3	TSST	Grip Strength	Angry Videos	Happy Videos	Treadmill Combo	
4	Angry Videos	Happy Videos	Treadmill Combo	TSST	Grip Strength	
5	Grip Strength	Angry Videos	Happy Videos	Treadmill Combo	TSST	
N	7	8	6	8	6	35

*After each task, participants were asked to rest calmly. Rest periods after each task are as follows: Angry Videos – 2 minutes; Grip Strength – 2 minutes; Happy Videos – 2 minutes; Treadmill Combo – 5 minutes; TSST – 2 minutes (Huang et al., 2010; Roth, Bachtler, & Fillingim, 1990; Rouselle, Blascovich, & Kelsey, 1995; Webb et al., 2011).

Independent Variables – Experimental task manipulations (Baseline, Treadmill Combo, Happiness-Inducing Videos, Anger-Inducing Videos, TSST, and Grip Strength) were utilized as independent variables in this experimental design.

Dependent Variables – Questionnaire data (PROMIS-Anger, PANAS-X, DASS 21, SUDS) were collected. Verbally reported SUDS ratings were utilized in the classifier modeling as well as physiological measures including PPG, EDA, and accelerometer data.

Data Analysis

The PPG data were sampled at 64 Hz, and the EDA signals were sampled at 4 Hz. Accelerometer data was sampled at 32 Hz. Event times, i.e., task boundaries, and Biopac (PPG) and (EDA) were stored in Biopac .acq files. Empatica E3 data (PPG, EDA, and accelerometer data) were stored in CSV files. All data was read into Python analysis scripts running under the Enthought Canopy environment where the numpy, scipy, pandas and matplotlib libraries were used for feature extraction and data analysis (McKinney, 2012), and the scikit-learn library (Garreta & Moncecchi, 2013) was used for classifier development.

Visual inspection of the raw data in the Biopac software and the interactive Python environment was used to discard bad (physiologically noisy or missing) participant data from either the Biopac or E3 sensor platforms, with good data (full and complete data without significant noise) available for 31 participants. Good data was sometimes available on the PPG or EDA channel, but not the other.

From the raw data, non-overlapping 1 minute windows were analyzed to yield feature vectors for the minute blocks. Inter-beat intervals (IBI) were extracted from the PPG data using a signal derivative-based algorithm based on

Johnston (2006), with modifications added to improve performance. For minutes with less than 40 valid IBI samples, the block of data was discarded but for other blocks the mean IBI was calculated. For each valid IBI block, the mean heart-rate (HR) was estimated by dividing the IBI mean from 60.0. For the EDA data, the mean was taken over the minute's raw data. For the accelerometer data, the magnitude of the x, y, z acceleration was calculated and the units were converted to units of g (gravitational acceleration constant). After 1 g was subtracted (the effect of gravitation on the E3 band), the values were rectified (i.e., absolute-value taken) and the mean was taken of this signal. This signal measures the amount of motion of the sensor band during the minute. The HR and EDA means were then normalized separately for each participant by subtracting the average of the 5 minute baseline task values for that participant from all of the values.

Matplotlib boxplots and scatterplots were used to explore the distributions of the task-specific patterns (e.g., Baseline and TSST-S) in feature space. Based on the distributions, the baseline-normalized HR and EDA means were used for stress vs. non-stress classification, and the motion measurement feature was used as input to a 1-feature Movement vs. Non-Movement classifier.

RESULTS

Experimental Stress Verification

In addition to the PPG and EDA data, the Subjective Unit of Distress Scale (SUDS) was collected. Following each segment, participants recorded on a scale from 0-100 their distress level during the phase. Figure 1 shows the distributions of stress ratings across all participants ($N = 35$). As Figure 1 shows, the TSST task led to the highest stress ratings, and the Baseline (initial 5 minute resting period) and Happy Videos tasks led to the lowest ratings. The effect of the task on the SUDS ratings was highly significant by a Kruskal-Wallis test ($H = 81.72, p < 0.001$).

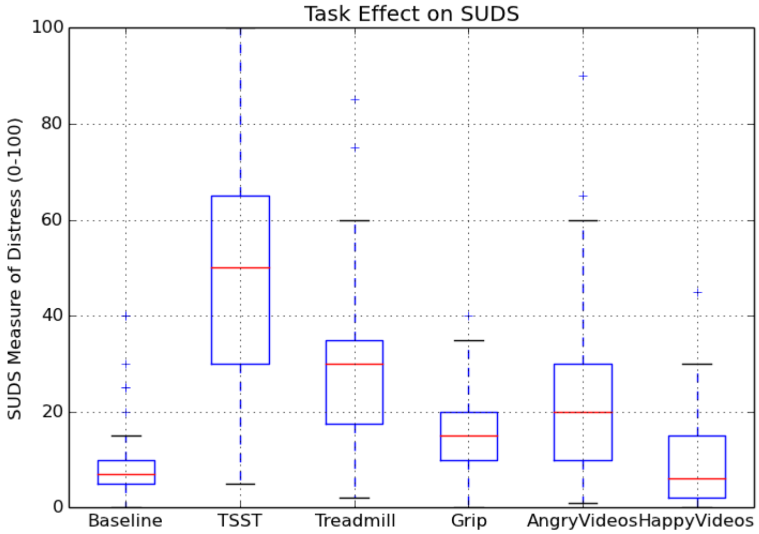


Figure 1. Effect of Task on Subjective Distress Measure

TSST induces the highest stress ratings, whereas baseline and happy videos induce the lowest. In Figure 1, red horizontal lines are medians; box ends are 25th and 75th percentiles. Whiskers are at a distance from the box ends of a minimum of 1.5 times the length of the box (interquartile range) and the data point at the extremum. Crosses are considered outliers.

Stress Classifier

The different task phases in the experiment were regarded as having distinct ground truth values for whether the participant would be considered stressed or not stressed. The speech (TSST-S) and mental arithmetic (TSST-A) phases of the TSST and the mental arithmetic component of the Treadmill task (TR-A) were considered to be psychological stress phases. The Baseline resting task (BASE), Treadmill-only task (TR), Happy Videos (HV), Grip task (GRIP), and any inter-task minutes (NULL) were considered to be occasions of non-psychological stress. For the preparatory phase of the TSST task (TSST-P) and the Angry Videos (AV) task, we made no assumption of stress vs. non-stress. Figure 2 shows the distributions of (non-normalized) heart rate (left) and skin conductance (right) data vs. task for all participants with E3 sensor data.

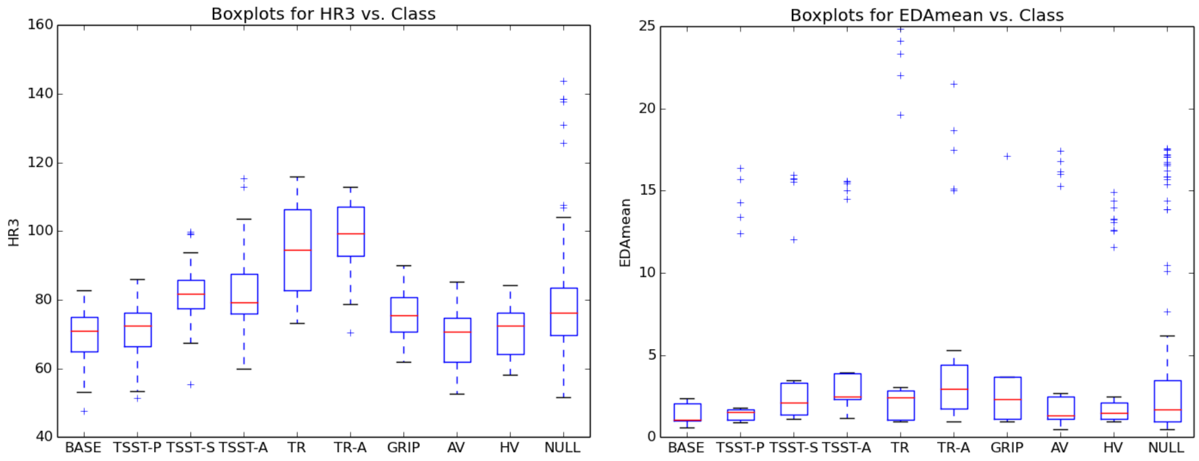


Figure 2. Effect of Task on Non-Normalized Heart Rate and Skin Conductance Measures

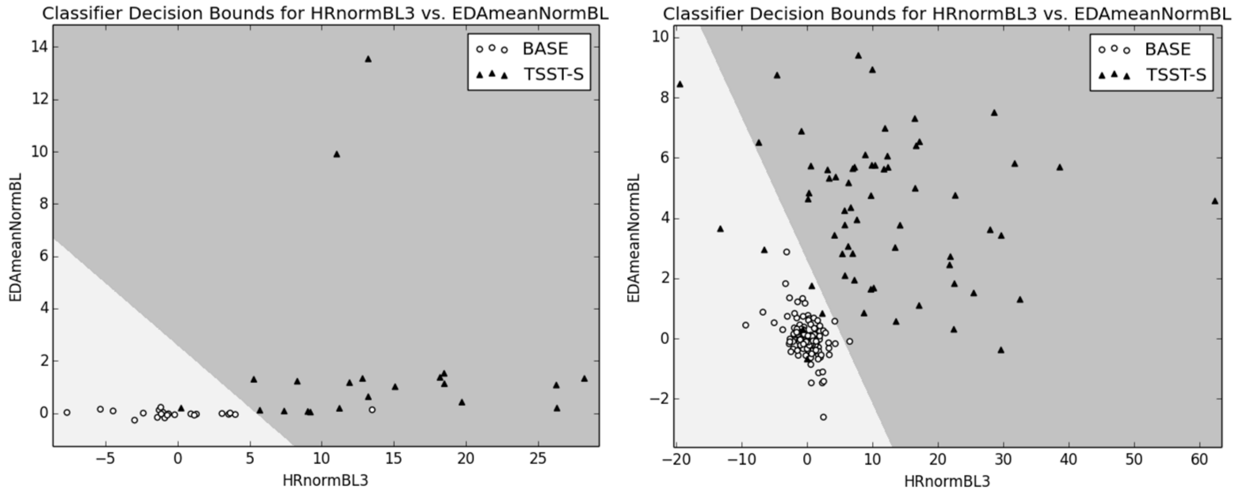
Figure 2 (left) includes heart-rate estimate distributions; TSST-S, TSST-A, TR, and TR-A have notably high HR distributions, whereas BASE tends to be low. Figure 3 (right) includes electrodermal activity estimate distributions; the BASE and HV conductances are relatively low, whereas TSST-S, TSST-A, TR, and TR-A are relatively high. Boxplots are arranged as in Figure 1.

HR and EDA tended to both be low during BASE. TSST-S, TSST-A, TR, and TR-A tended to be relatively high. The fact that both TR and TR-A tended to have high physiological responses—for HR, higher than the TSST phases— illustrates the difficulty at avoiding confounds between stress and physical activity.

In the end, we opted to train a psychological stress vs. non-stress classifier using a “best vs. worst” case classification of BASE vs. TSST-S, using the Baseline-normalized mean HR and EDA features. Subtracting the baseline means for each participant allowed an improvement in the performance for a subject-independent classifier. We settled on a 2-feature linear model trained with stochastic gradient descent. This architecture has the advantage of having a bias term that may be used to tune the decision boundary threshold on the stress vs. non-stress classifier to allow adjusting the tradeoff between hit and false alarm rates, potentially setting a different threshold for each participant, which could improve performance of the classifier significantly for the individual patient-users of the stress detection system.

With the stochastic gradient descent learning on a linear model, we began by using 5-fold cross-validation to evaluate the average performance of the algorithm. The train / test set consisted of E3 data feature vectors taken from the BASE and TSST-S minutes of the participants who had both good HR and EDA data. For the cross-validation, this set was divided into fifths and, iteratively, 1 of the 5 blocks was left out for testing and the other 4 were used to train the classifier. The cross-validation score was achieved by averaging together the test-set classification accuracy scores; for the E3 classifier, this score was 85.7%.

Figure 3 shows the classification results of training a psychological stress vs. non-stress classifier using 75% of the E3 physiological data which had both good HR and EDA data. The left panel shows the learned stress vs. non-stress decision boundary and how the train and test E3 data from the participants with good HR and EDA data lines up with this boundary. Training accuracy is a classifier performance metric defined by signal detection theory (Green & Swets, 1966): (hit = classifier correctly identified spike in negative arousal, miss = classifier missed a spike in negative arousal, false alarm = classifier identified a spike in negative arousal when one did not actually occur, and correct rejection = classifier did not identify a spike in negative arousal when one did not occur). The training accuracy was 97.1%. Test set accuracy on the remaining 25% of the data was 91.7% (11 / 12). The hit rate on the test set was 100% (4 / 4), and the false-alarm rate was 12.5% (1 / 8). It turns out that the model trained on the E3 data also performed well in classifying the larger Biopac-collected data set (Figure 3(right)). 25 of the participants had both good HR and EDA data and the accuracy of the classifier in distinguishing all BASE vs. TSST-S minutes was 95.1% (176 / 185). The hit rate was 89.1% (57 / 64) and the false alarm rate 1.7% (2 / 121). Of the two panels of Figure 3, it should be noted that, although the slopes of the decision boundaries appear different, they are in fact the same when axes are identically scaled; the Biopac data has a much wider spread in the normalized EDA dimension, but the data still falls within the decision boundary learned from the E3 data.


Classifier Decision Bounds for HRnormBL3 vs. EDAmeanNormBL

Classifier Decision Bounds for HRnormBL3 vs. EDAmeanNormBL

Figure 3. Decision Boundary for Linear Stochastic Gradient Descent- Trained E3 Classifier Applied to E3 (left) and Biopac (right) Data

A linear classifier was trained using stochastic gradient descent learning on 75% of the BASE and TSST-S E3 data where the features were Baseline-normalized mean HR and EDA. (The entire train / test set is shown in Figure 3). The circular dots show the location of BASE physiological responses in feature space; the triangular dots show the location of TSST-S responses. The light grey-shaded portion of the graph shows the part of the feature space which is classified as Non-Stress, whereas the dark grey-shaded portion shows the Stress-classified region. The same trained E3 model, surprisingly, is able to perform well at classifying most of the Biopac physiological data (25 participants that had both good HR and EDA data). Note that the decision boundaries in both panels are identical if the axes are scaled to be equivalent.

Movement Detection

The classifier shown in Figure 3 can be used as a provisional psychological stress vs. non-stress classifier, but one remaining issue is that it can lead to false alarm psychological stress classifications during minutes when there is a high amount of physical stress, especially during the treadmill task. Our solution for this problem is to detect minutes when there is high movement and deactivate the stress vs. non-stress classifier during those minutes, as there is too great a risk of confounding results using that classifier during moments of high physical activity.

Figure 4 shows how the trial task affects the motion-measure. Both treadmill tasks (TR and TR-A) have significantly higher responses of this feature, clearly allowing the high movement tasks to be separated from the rest. It should be noted that some of the non-task (NULL) minutes manifest high motion artifact, but this is likely to be because participants had to often walk between experimental stations between task phases.

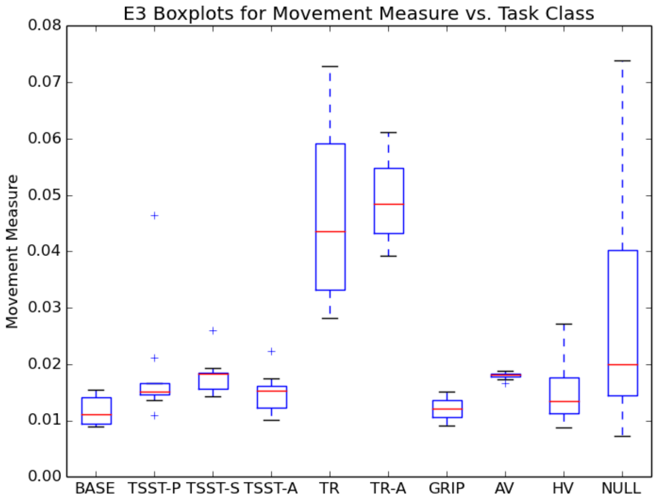


Figure 4. Effect of Task on Motion Measure

The treadmill tasks (TR, TR-A) have significantly more activity than the others (see Figure 4). The non-task (NULL) minutes have some high motion measurements, but these can be explained by the fact that participants between tasks often had to walk to the next experimental station.

Motion estimates were collected using accelerometer data, and a single feature was then used to train a movement vs. non-movement classifier, again with a linear classifier trained with stochastic gradient descent. TR and TR-A minutes were labeled as movement minutes. BASE, TSST-P, TSST-S, TSST-A, GRIP, AV, and HV minutes were labeled as non-movement. The non-task (NULL) minutes were not included in training and testing. 5-fold cross-validation gave an average accuracy of 97.9%. Training accuracy with 75% of the data led to a 98.0% accuracy. Testing on the remaining 25% of the data led to an accuracy of 97.0% (32 / 33). On this same test set, the hit rate was 100% (8 / 8) and the false alarm rate was 4% (1 / 25). Such a classifier is likely to be able to eliminate high physical activity minutes from consideration by the stress detector, which should reduce the false alarm rates.

DISCUSSION

Although the main objective of developing a mobile classifier using non-invasive sensors to classify psychological stress with over 85% accuracy was achieved, further refinement is needed to maintain the high level of accuracy across a variety of users and environmental conditions.

Future research will include further accuracy refining through reduction in environmental noise, a smart algorithm to learn individual user stress thresholds, and differentiation among different types of arousal. Extensive operational testing to reduce environmental noise is being conducted to determine the changes in classifier false alarms and misses when collecting data in different temperatures and while performing different physical activities from typing on a keyboard to walking or running. In addition, a variety of users will undergo operational testing to determine if the accuracy changes based on user demographics such as age and gender. Initial findings have confirmed past literature citing the improved accuracy with personalized classifiers (Alamudun et al., 2012; de Santos Sierra et al., 2011), and plans to refine the algorithm to adjust alert thresholds to each user are being developed. This smart version of the algorithm will automatically “learn” each user’s true stress thresholds to prevent high false alarms or misses and improve accuracy. Finally, the algorithm will be refined to differentiate between negative arousal and positive arousal, which are very different in semantics but physiologically appear very similar. The planned refinements to this algorithm are necessary to maintaining a high level of accuracy while being a practical solution for real world applications

Applications

The primary intended application for the algorithm developed under this research is to be integrated in a mobile software package to support stress therapy for Veterans. This system collects and processes the real-time objective stress data from the wrist-worn sensor suite to alert the user of heightened stress via a mobile application. The mobile application also includes subjective stress data collection and relaxation support. This data is available to the treatment provider via a secure web-based portal to better understand trends throughout the stress therapy treatment process (https://mcalm.designinteractive.net). The algorithm resulting from this research will be evaluated in a clinical environment and integrated in a mobile software package.

The capability to monitor high quality, unobtrusive, real-time physiological data in a mobile environment has additional opportunities outside of Veteran mental health treatment, including military training. Military training instructors could remotely and simultaneously monitor objective psychological stress status for each trainee during live training sessions and act on the information in real-time. For example, an instructor could amplify environmental stressors such as auditory noise or olfactory cues when trainees are operating at low arousal levels. Or, an instructor can identify individual trainees that tend to have more intense stress responses than others during training scenarios to provide targeted coping and resilience training interventions. Not only does this capability support live stress training, but also the evaluation of training systems designed to induce stress through simulations.

Beyond military training applications, additional applications for this capability include stress research for laboratory and field settings, deception detection in mobile environments with minimal equipment required, chronic disease monitoring for tracking outpatient health and long-term data capture to inform care, and objective, real-time user experience evaluations. Sensor-based and remote monitoring technologies such as the capability described in this research will soon transform the way we care for ourselves and loved ones in our everyday lives. Mobile health applications and wearable physiological sensors have the potential to analyze and present meaningful data to better manage and optimize general health and specific health conditions, however, robust algorithms such as the one presented in this paper are a necessary component to realizing the potential of this technology.

REFERENCES

Alamudun, F., Choi, J., Gutierrez-Osuna, R., Khan, H., & Ahmed, B. (2012). Removal of subject-dependent and activity-dependent variation in physiological measures of stress. In *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2012 6th International Conference on* (pp. 115-122). IEEE.

Bakker, J., Pechenizkiy, M., & Sidorova, N. (2011). What's your current stress level? Detection of stress patterns from GSR sensor data. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on* (pp. 573-580). IEEE.

Bartolini, E. E. (2011). Eliciting emotion with film: Development of a stimulus set (Master's Thesis). Retrieved fromhttp://wesscholar.wesleyan.edu/cgi/viewcontent.cgi?article=1616&context=etd_hon_theses

Bartone, P. (2006). Resilience Under Military Operational Stress: Can Leaders Influence Hardiness? Military Psychology; 18: 131-148.

Carroll, M., Hale, K., Stanney, K., Woodman, M., DeVore, L., Squire, P., & Sciarini, L. (2012). Framework for Training Adaptable and Stress-Resilient Decision Making. Interservice Training, Simulation, and Education Conference (I/ITSEC) 2012.

Choi, J., Ahmed, B., & Gutierrez-Osuna, R. (2012). Development and evaluation of an ambulatory stress monitor based on wearable sensors. *Information Technology in Biomedicine, IEEE Transactions on*, 16(2), 279-286.

Dickerson, S. S. & Kemeny, M. E. (2004). Acute stressors and cortisol responses: A theoretical integration and synthesis of laboratory research. *Psychological Bulletin, 130*(3), 355-391.

Everly, G. S., & Lating, J. M. (2013). The Anatomy and Physiology of the Human Stress Response. *A clinical guide to the treatment of the human stress response*, 17-51.

Follin, A. (1994). Preventing post-traumatic stress disorder resulting from military operations. Military Medicine, 159(12); 739-746.

Garreta, R., & Moncecchi, G. (2013). *Learning Scikit-learn: Machine Learning in Python*. Packt Publishing Ltd.

Green, D. & Swets, J. (1966). Signal Detection Theory and Psychophysics. Wiley & Sons, Inc., New York

Johnston, W. S. (2006). *Development of a Signal Processing Library for Extraction of SpO2, HR, HRV, and RR from Photoplethysmographic Waveforms* (Doctoral dissertation, WORCESTER POLYTECHNIC INSTITUTE).

Huang, C., Webb, H. E., Garten, R. S., Kamimori, G. H., Evans, R. K., & Acevedo, E. O. (2010). Stress hormones and immunological responses to a dual challenge in professional firefighters. *International Journal of Psychophysiology, 75*(2010), 312-318.

Jenkins, L. M. & Andewes, D. G. (2012). A new set of standardised verbal and non-verbal contemporary film stimuli for the elicitation of emotions. *Brain Impairment, 13*(2), 212-227.

Johnston, W. S. (2006). *Development of a Signal Processing Library for Extraction of SpO2, HR, HRV, and RR from Photoplethysmographic Waveforms* (Doctoral dissertation, WORCESTER POLYTECHNIC INSTITUTE).

Jones, D., Hale, K., Dechmerowski, S., & Fouad, H. (2012). Creating Adaptive Emotional Experience During VE Training. Interservice Training, Simulation, and Education Conference (I/ITSEC) 2012.

Kavanagh, J. (2005). Stress and Performance: A Review of the Literature and its Applicability to the Military. RAND technical report.

Kirschbaum, C., Pirke, K., & Hellhammer, D. H. (1993). The 'Trier Social Stress Test' – a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology, 28*, 76-81.

Malik, M., Bigger, J. T., Camm, A. J., Kleiger, R. E., Malliani, A., Moss, A. J., & Schwartz, P. J. (1996). Heart rate variability standards of measurement, physiological interpretation, and clinical use. *European heart journal*, 17(3), 354-381.

Marshall, R. D., Olfson, M., Hellman, F., Blanco, C., & Struening, E. L. (2001). Comorbidity, impairment, and suicidality in subthreshold PTSD. *American Journal of Psychiatry, 158*(9), 1467-1473. McKinney, W. (2012). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. O'Reilly Media.

National Alliance on Mental Illness [NAMI] (2014). Retrieved February 21, 2014 from http://www.nami.org/template.cfm?section=mental_illnesses1

Pitman, R K., Rasmusson, A.M., Koenen, K.C., Shin, L.M., Orr, S.P., Gilbertson, M.W., & Liberzon, I. (2012). Biological studies of post-traumatic stress disorder. *Nature Reviews Neuroscience*, *13*(11), 769-787.

Plarre, K., Raij, A., Hossain, S. M., Ali, A. A., Nakajima, M., al' Absi, M., Ertin, E., Kamarck, T., Kumar, S., Scott, M., Siewiorek, D., Smailagic, A., & Wittmers, Jr., L. E. (2011). Continuous Inference of Psychological Stress from Sensory Measurements Collected in the Natural Environment. In *Information Processing in Sensor Networks (IPSN), 2011 10th International Conference on* (pp. 97-108). IEEE.

RAND (2008). Invisible wounds of war : psychological and cognitive injuries, their consequences, and services to assist recovery. Terri Tanielian, Lisa H. Jaycox (eds).

Rousselle, J. G., Blascovich, J., & Kelsey, R. M. (1995). Cardiorespiratory response under combined psychological and exercise stress.*International Journal of Psychophysiology*, *20*(1995), 49-58.

Roth, D. L., Bachtler, S. D., & Fillingim, R. B. (1990). Acute emotional and cardiovascular effects of stressful mental work during aerobic exercise. *Psychophysiology*, *27*(6), 694-701.

Russo, S. J., Murrough, J.W., Han, M.H., Charney, D.S. & Nestler, E.J. (2012). Neurobiology of resilience. *Nature Neuroscience*, 15(11): 1475-1484.

de Santos Sierra, A., Ávila, C. S., Casanova, J. G., & del Pozo, G. B. (2011). Real-Time Stress Detection by Means of Physiological Signals. *Group of Biometrics, Biosignals and Security Universidad Politécnica de Madrid, Spain*, 24-44.

Saunders, T., Driskell, J., Hall, J., & Salas, E. (1996). The Effect of Stress Inoculation Training on Anxiety and Performance. ARI Research Note 96-27.

Schaefer, A., Nils, F., Sanchez, X., & Philippot, P. (2010). Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers. *Cognition and Emotion, 24*(7), 1153-1172.

Sun, F. T., Kuo, C., Cheng, H. T., Buthpitiya, S., Collins, P., & Griss, M. (2012). Activity-aware mental stress detection using physiological sensors. In *Mobile Computing, Applications, and Services* (pp. 211-230). Springer Berlin Heidelberg.

Vijayalakshmi, P., Madanmohan, Bhavanani, A. B., Patil, A., & Babu, K. P. (2004). Modulation of stress induced by isometric handgrip test in hypertensive patients following yogic relaxation training. *Indian Journal of Physiological Pharmacology, 48*(1), 59-64.

Webb, H. E., Fabianke-Kadue, E. C., Kraemer, R. R., Kamimori, G. H., Castracane, V. D., & Acevedo, E. O. (2011). Stress reactivity to repeated low-level challenges: A pilot study. *Applied Psychophysiology Biofeedback, 36*, 243-250.

Yarvis, J. S. & Schiess, L. (2008). Subthreshold PTSD as a predictor of depression, alcohol use, and health problems in soldiers. *Journal of Workplace Behavioral Health, 23*(4), 395-424.