The Case for a Commodity Hardware Solution for Stress Detection

Varun Mishra
Dartmouth College
varun@cs.dartmouth.edu

Gunnar Pope
Dartmouth College

Sarah Lord
Dartmouth College

Stephanie Lewia
Dartmouth College

Byron Lowens
Clemson University

Kelly Caine
Clemson University

Sougata Sen
Dartmouth College

Ryan Halter
Dartmouth College

David Kotz
Dartmouth College

ABSTRACT

Timely detection of an individual's stress level has the potential to expedite and improve stress management, thereby reducing the risk of adverse health consequences that may arise due to unawareness or mismanagement of stress. Recent advances in wearable sensing have resulted in multiple approaches to detect and monitor stress with varying levels of accuracy. The most accurate methods, however, rely on clinical grade sensors strapped to the user. These sensors measure physiological signals of a person

UbiComp/ISWC'18 Adjunct, October 8–12, 2018, Singapore, Singapore

ACM ISBN 978-1-4503-5966-5/18/10…$15.00
https://doi.org/10.1145/3267305.3267538

and are often bulky, custom-made, expensive, and/or in limited supply, hence limiting their large-scale adoption by researchers and the general public. In this paper, we explore the viability of commercially available off-the-shelf sensors for stress monitoring. The idea is to be able to use cheap, non-clinical sensors to capture physiological signals, and make inferences about the wearer's stress level based on that data. In this paper, we describe a system involving a popular off-the-shelf heart-rate monitor, the Polar H7; we evaluated our system in a lab setting with three well-validated stress-inducing stimuli with 26 participants. Our analysis shows that using the off-the-shelf sensor alone, we were able to detect stressful events with an F1 score of 0.81, on par with clinical-grade sensors.

CCS CONCEPTS

• **Human-centered computing** → **Ubiquitous and mobile computing**; • **Applied computing** → *Health care information systems*; *Health informatics*;

KEYWORDS

Stress detection, mobile health (mHealth), commodity wearables, mental health

ACM Reference Format:
Varun Mishra, Gunnar Pope, Sarah Lord, Stephanie Lewia, Byron Lowens, Kelly Caine, Sougata Sen, Ryan Halter, and David Kotz. 2018. The Case for a Commodity Hardware Solution for Stress Detection. In *Adjunct Proceedings of the 2018 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2018 International Symposium on Wearable Computers (UbiComp/ISWC'18 Adjunct), October 8–12, 2018, Singapore, Singapore.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3267305.3267538

INTRODUCTION

Stress is defined as the brain's response to any demand or change in the external environment [15], and has the potential to actuate changes within an individual's lifestyle. When an individual experiences sustained stress over long periods of time, it could lead to *chronic stress*, which is severely detrimental to both physical and mental health [1]. Hence, timely detection and intervention is extremely important.

With recent advancements in sensor and wearable technologies, researchers are beginning to make progress on continuous and passive measurement of stress [8, 10–12, 16, 18]. While this prior work introduces and studies a variety of wearable devices and sensors to capture physiological data with a focus on detecting or predicting stress (or stressful events), it relies on custom-made, or clinical-grade sensors, which are often bulky, uncomfortable, inaccessible and/or expensive, making them unappealing or out of reach for many (a brief summary is provided in Sidebar 1). These limitations have prevented large-scale adoption of such sensors by (a) researchers who want to observe participant stress in real or near-real time, (b) researchers who want to study interventions and their effect on other behaviors such as anxiety, smoking cessation or drug abuse, and (c) consumers who want to

In several past work, the researchers developed their own custom-fitted sensing system [4, 10, 11, 16, 18]. While the benefits of using a custom sensor-suite may include – higher-quality signals, control over signal type/frequency, control over battery life, and so forth, they also have some major limitations, i.e., lack of reproducibility by other researchers, lack of large-scale deployments, and unavailability to other researchers who want to use similar sensors for detecting other health outcomes.

There are some work that have used a *commercially* available sensor/device. Muaremi et al. used a combination of the Zephyr BioHarness 3.0 [22] and an Empatica E3 [7] for monitoring stress while sleeping [14]. Gjoreski et al. used the Empatica E3 to detect stress in a lab and an unconstrained field (free-living) setting [8]. These devices, however, are very expensive (the Zephyr BioHarness sells on Amazon for over $650, Empatica E3 has been replaced with E4, which sells for over $1600 and has a 2–6 week shipping time). The high cost of these sensors limit large-scale deployments of these devices in studies not only for stress detection, but also for other mental and behavioral health outcomes.

In contrast, we use a *commodity* device, the Polar H7 heart-rate monitor [17], which is available on Amazon for just under $70, and has been recently updated with the newer Polar H10 which is available for $89.

Sidebar 1: Summary of Related Work

Study Description
Participants: $n = 26$ participants (14 females, 12 males; 12 undergraduate and 14 graduate students), with a mean age of 23 ± 3.24 years.
Data Types: Heart-rate and R-R intervals.
Device used: Polar H7 [17]. We conducted a preliminary test to choose between the Polar H7 and the Zephyr HXM. We compared both these devices alongside a clinical ECG device, the Biopac MP150 [3], and found that as compared to Zephyr HXM, the features computed from Polar H7 had a stronger correlation with those computed from the Biopac.
Lab Protocol: After signing the consent form, the participants experienced three types of stressors – mental arithmetic, startle response, and cold water – all well-validated stimuli known to induce stress.

Specifically, the protocol was:

(1) Resting baseline – participant sat in a resting position for 10 minutes.
(2) Mental arithmetic task – participant counted backwards in steps of 7 (4 minutes).
(3) Rest period – participant sat in a resting position for 5 minutes, to allow him/her to return to baseline.
(4) Startle response test – participant faced away from the lab staff and closed his/her eyes; staff then dropped a book at an several random and unexpected moments, startling the participant (4 minutes).
(5) Rest period – 5 minutes, as before.
(6) Cold water test – participant submerged his/her right hand in a bucket of ice water for as long as tolerable (up to 4 minutes).
(7) Rest period – 5 minutes, as before.

Sidebar 2: The data collection protocol during our study.

monitor their stress level beyond the clinical setting, in free-living conditions. In this work, we aim to answer the following question: *Can a commodity device be used to accurately measure stress?*

To answer this question, we conducted a study with 26 participants using an *off-the-shelf* commodity device heart-rate monitor (Polar H7 [17]). Our analysis shows that the features computed from data collected by a commodity heart-rate sensor show statistically significant differences between baseline *rest* and *stressed* periods. Furthermore, the features also show significant difference across several types of stress-inducing stimuli.

Further, by performing a thorough evaluation in the lab setting (18 hours of data), we demonstrate methods for accurately detecting stress with an F1 score of 0.81. The results are on-par with results attained in prior research that uses clinical-grade Electrocardiography (ECG) sensors to identify stressful periods in the lab [11].

These results give us confidence about the usability of commodity heart-rate monitors (in this case the Polar H7). While more analyses with a varied user-base is required, we believe this is a strong step in the direction of eliminating researchers' dependence on custom or expensive clinical-grade ECG monitors for stress measurement, and possibly to other mental and behavioral health outcomes.

DATA COLLECTION AND PROCESSING

We now discuss our data collection, followed by methods for processing the data, which includes data cleaning, normalization, and feature computation and selection.

Data Collection

We conducted a study comprising lab and field components. All participants completed both the lab and field components and were compensated with $50 for their time. The analysis and results discussed in this work is *only* from the lab component of the study; the data analysis from the field component is currently underway. The study description, along with the lab protocol, is explained in Sidebar 2. At the end of the initial baseline rest period and after each stressor, we asked the participant to verbally rate their stress level on a scale of 1–5; this was the stress perceived by the user. As the ground truth, we labeled each minute of data collected in the lab as *stressed* (class = 1) or *not stressed* (class = 0), based on whether the participant was experiencing a stressor stimulus within that minute. This study was approved by our Institutional Review Board (IRB).

Data Cleaning

We begin with preliminary data cleaning, to filter out invalid data points. In this step we are not trying to handle outliers (which may or may not be valid readings), but remove obviously erroneous data readings. This step is important because the sensors used for physiological measurements are not clinical quality, and may need a few seconds to acquire stable heart-rate readings. We noticed

these erroneous readings usually occurred when a participant was trying to put on or remove the device, or when the device did not snugly fit the participant.

If the heart-rate value was outside a pre-determined range, we dropped both the heart-rate value and any R-R interval (i.e., the time interval between two consecutive R peaks in the QRS complex of an ECG wave. R-R interval is a measure of inter-beat variability, also known as Heart-rate Variability (HRV), and has been shown to be a marker for stress and health.) values received in that second. Based on previous research conducted to find the maximum human heart rate [6, 20], we set our upper bound to 220 bpm. To determine the lower bound, we inspected heart-rate data of all the participants (visually) to find any noticeable value that would seem invalid. The resulting range [30:220] bpm is very conservative; we are confident that any data point outside this range is invalid.

Feature Computation

We next use the data remaining after the previous steps to compute features to quantify heart-rate variability (HRV). We split the data into one-minute intervals, and compute a set of features for each interval. However, before we compute some features for further analyses, it is critical that we (1) handle the effect of outliers in the data and (2) remove any participant-specific effects on the data, so as to create a generalized model, without any participant dependency. These issues would significantly impact the computed features, and eventually the accuracy of the results obtained. We thus look at each in more detail to understand how the results change with different methods for handling outliers and normalization. All of the previous works we reviewed seem to have just selected some method for handling outliers (if any) and normalization, without taking into account the effect of their choice on the outcome of the metrics under study.

Outliers. While dealing with outliers in data, the common approaches are (a) leave them in the data, (b) reduce the effect the outliers might have, or (c) remove them completely. In our work we look at each of these approaches and their effect on model training and evaluation. For the first approach, we do nothing to the data, i.e., leave it as-is. In the second approach, we use *winsorization* to reduce the effect of outliers on the dataset [21]. This approach was also used by some previous work, e.g., cStress [11] and Gjoreski et al. [9]. For the third approach (c), we simply remove (trim) data points that we deem as outliers.

We define *outlier* as a point that lies beyond a certain threshold above or below the median of the data. For our purposes we choose the threshold as three times the median absolute deviation (MAD) within that participant's data. This choice ensures that we considered only the extreme values as outliers, and over 99% of the data is unaltered. Having defined *outlier*, we establish the bounds as $median \pm 3 \times MAD$

	Outliers present	*Winsorization*	*Trimming*
Minmax	outlier_minmax	wins_minmax	trim_minmax
z-score	outlier_zscore	wins_zscore	trim_zscore

Table 1: Nomenclature for the different combinations of outlier handling and normalization methods.

Heart rate	*R-R interval*
mean, median, standard deviation, 80th percentile, 20th percentile	mean, median, standard deviation, max, min, 80th percentile, 20th percentile, root mean square of successive differences (RMSSD)

Table 2: All the features computed from the filtered and normalized HRV data segregated by the base measures – heart-rate, and R-R interval.

The next steps are straightforward; when winsorizing, we replace any value greater than the *upper_bound* with the *upper_bound* value, and any value lesser than the *lower_bound* with the *lower_bound* value. Alternately, for trimming we just drop the values less than the *lower_bound* or greater than the *upper_bound*.

It is important to note that handling outliers by both winsorization and trimming is done individually for each participant.

Normalization. is important to remove participant-specific effects on the data, so as to make the model generalizable to any participant. We tried two different methods for data normalization. With physiological data (e.g., heart rate, Galvanic Skin Response (GSR), skin temperature) each participant has a different natural range. Hence, the first normalization method we try is *minmax* normalization, which simply transforms the values into the range $[0, 1]$. Given a vector $x = (x_1, x_2, \ldots, x_n)$, the *minmax* normalized value for the i^{th} element in x is given by,

$$z_i = \frac{x_i - min(x)}{max(x) - min(x)}$$

Further, there might be more intrinsic participant effects – participant-specific mean and standard deviation, hence as the second normalization technique, we tried z-score normalization. In case of z-score normalization, the normalized value z_i is denoted by,

$$z_i = \frac{x_i - \mu}{\sigma}$$

where, μ is the mean of x and σ is the standard deviation of x. It would be interesting to observe the role participant-specific effects have on model training and validation. We go through both of the normalization steps individually for all three ways of handling outliers. Table 1 provides our nomenclature for each of the methods we used.

Feature computation. We grouped the normalized data into one-minute windows. Given the short duration of our lab experiments, we wanted to select the shortest possible window size. Esco and Flatt demonstrated that as compared to 10- or 30-second windows, the features computed in the 60-second window size had the highest agreement with the conventional 5-minute window size [5]. Furthermore, the one-minute window has been common in physiological monitoring [10, 11, 16].

For the HRV data, we selected only the time-domain features for our work, as shown in Table 2. All of these time-domain features have been shown to be effective in predicting stressful periods by other researchers [11]. Unlike earlier work, however, we avoided frequency-domain features (e.g., low-frequency (LF) bands, high-frequency (HF) bands, and low:high frequency (LF:HF) ratio) for the following reasons.

Features	Math Test		Book Test		Cold Test	
	t-stat	*p-value*	*t-stat*	*p-value*	*t-stat*	*p-value*
mean HR	**−14.170**	**<0.001**	**7.490**	**<0.001**	−1.420	0.159
standard deviation HR	−0.560	0.579	−0.810	0.419	−1.300	0.198
median HR	**−13.970**	**<0.001**	**7.670**	**<0.001**	−1.240	0.217
20th percentile HR	**−12.540**	**<0.001**	**7.710**	**<0.001**	−0.930	0.355
80th percentile HR	**−13.750**	**<0.001**	**6.380**	**<0.001**	−1.770	0.080
mean R-R	**7.020**	**<0.001**	**−5.830**	**<0.001**	−0.770	0.443
standard deviation R-R	0.220	0.830	−0.350	0.726	1.140	0.254
median R-R	**6.870**	**<0.001**	**−6.760**	**<0.001**	−0.380	0.704
max R-R	**6.790**	**<0.001**	**−6.740**	**<0.001**	0.200	0.843
min R-R	**2.650**	**0.009**	0.180	0.858	0.680	0.496
20th percentile R-R	**3.630**	**<0.001**	**−3.270**	**0.001**	−1.780	0.076
80th percentile R-R	**10.680**	**<0.001**	**−7.860**	**<0.001**	0.180	0.856
RMSSD	−0.470	0.637	−0.780	0.436	0.300	0.765

Table 3: Significant heart-rate based feature differences from initial rest period of 10 minutes. Significant scores ($p < 0.05$) are shown in bold.

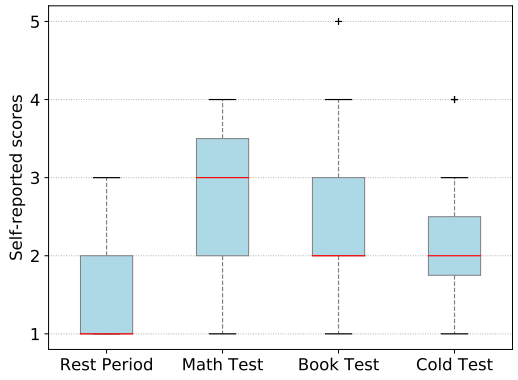


Figure 1: Participant self-reports after each lab period.

The RMSSD (root mean square of successive differences of successive R-R intervals) is associated to short-term changes in the heart, and is considered to be a solid measure of vagal tone and parasympathetic activity, similar to HF [13]. Several studies have also shown that RMSSD and HF are highly correlated [19]. Further, unlike HF, RMSSD is easier to compute, and is not affected by other confounding factors such as breathing. Hence, we felt RMSSD a good alternative to HF, thereby nullifying the need to compute HF.

Unlike HF, which represents parasympathetic activity, LF is less clear. While some researchers believe LF represents sympathetic activity, others suggest it is a mix of both sympathetic and parasympathetic activities [2]. Furthermore, the rationale behind using LF:HF ratio is that since HF represents parasympathetic activity, a lower HF will increase the ratio, suggesting more stress; but, since the role of LF is not really clear, looking at the ratio might be misleading as well [2]. Also, for computing LF, we need a window size of at least 2 minutes, which would reduce our data size by half. Furthermore, earlier work like *cStress* found that compared to other time-domain features, and HF, the feature importance of LF and LF:HF is extremely low [11]. Hence we decided to leave out LF and LF:HF features from our work, thus not requiring us to calculate any frequency-domain features.

EVALUATION

In this section we evaluate our approach. We begin by determining whether we were able to capture a significant difference between the *resting* and *stress-induced* periods of the lab component, followed by building and evaluating machine-learning models from the lab dataset, and finally using the models built in the lab to infer *stress/not-stress* in the field.

Significant features

We first determined whether we could distinguish between resting state and stressful states, in the lab data. To this end, we use features computed from the first 10 minutes of the initial rest period and compare them individually to the features computed from the Math Test, Book Test, and Cold Test respectively. We used Welsh's t-test of unequal variances to determine which features showed any statistically significant differences between the resting baseline period and each of the stress-induction periods. As described above, we followed three ways of handling outliers and two ways for data normalization, leading to a total of six combinations, as shown in Table 1. Across all the six combinations, we observed the maximum number of features showing significant differences in the *trim_zscore* combination, and for the sake of space, we report results only for that one combination, i.e., trimmed outliers and z-score normalization.

The results for the heart-rate features are shown in Table 3. It is evident that for the Math Test and the Book Test, there are several features that showed statistically significant differences. This is, however, not the case for the Cold Test, where we found no feature showing statistically significant

Features	Math Test		Book Test		Cold Test	
	t-stat	*p-value*	*t-stat*	*p-value*	*t-stat*	*p-value*
mean HR	**−13.230**	**<0.001**	**5.970**	**<0.001**	−1.740	0.084
standard deviation HR	−1.270	0.204	−1.530	0.129	−1.710	0.090
median HR	**−13.060**	**<0.001**	**6.080**	**<0.001**	−1.620	0.107
20th percentile HR	**−11.640**	**<0.001**	**6.240**	**<0.001**	−1.160	0.249
80th percentile HR	**−12.810**	**<0.001**	**5.030**	**<0.001**	**−2.040**	**0.044**
mean R-R	**13.920**	**<0.001**	**−6.380**	**<0.001**	**2.110**	**0.037**
standard deviation R-R	0.350	0.725	−2.440	0.015	−0.790	0.428
median R-R	**14.150**	**<0.001**	**−6.250**	**<0.001**	1.970	0.052
max R-R	**6.760**	**<0.001**	**−4.960**	**<0.001**	1.130	0.262
min R-R	**8.730**	**<0.001**	−1.760	0.080	**2.460**	**0.015**
20th percentile R-R	**13.440**	**<0.001**	**−4.420**	**<0.001**	**2.410**	**0.017**
80th percentile R-R	**11.160**	**<0.001**	**−6.430**	**<0.001**	1.210	0.228
RMSSD	0.420	0.676	**−2.670**	**0.008**	−1.200	0.231

Table 4: Significant heart-rate based feature differences from the last 4 minutes of the initial rest period. Significant scores ($p < 0.05$) are shown in bold.

	trim_zscore		*trim_minmax*		*wins_zscore*		*wins_minmax*	
Metrics	SVM	RF	SVM	RF	SVM	RF	SVM	RF
Precision	0.64	0.62	0.60	0.66	0.68	0.61	0.61	0.62
Recall	0.72	0.66	0.52	0.70	0.59	0.66	0.48	0.68
F1 score	0.68	0.64	0.56	0.68	0.63	0.63	0.53	0.65

Table 5: LOSO Cross-validation results from the different datasets, using SVM and RF, and considering the entire rest baseline of 10 minutes as *not stressed*

	trim_zscore		*trim_minmax*		*wins_zscore*		*wins_minmax*	
Metrics	SVM	RF	SVM	RF	SVM	RF	SVM	RF
Precision	0.80	0.78	0.70	0.81	0.79	0.78	0.76	0.78
Recall	0.81	0.70	0.59	0.67	0.69	0.68	0.59	0.67
F1 score	0.81	0.73	0.69	0.73	0.73	0.72	0.66	0.72

Table 6: LOSO Cross-validation results from the different datasets, using SVM and RF, and considering only the last 4 minutes of the rest baseline as *not stressed*

difference from the initial 10-minute rest baseline. This result was unexpected, which suggested that the Cold Test was not affecting (i.e., stressing) the participants significantly from the baseline resting period. This result prompted us to look at the self-reports the participants answered (on a scale of 1 to 5), during the lab study, after the baseline rest period, and after each of the stress tests, as shown in Figure 1. We observed that most participants gave a lower stress score after the Cold Test, as compared to the previous two tests.

A two-tailed unpaired t-test between the self-reported scores after the baseline rest period and the Cold Test across all participants, however, revealed a statistically significant difference: $t\_stat = 3.4734; p = 0.001$.

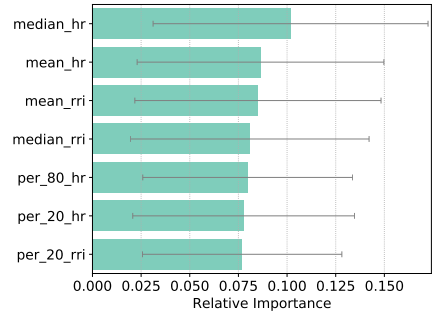
Due to this significant difference between the participants' responses, we hypothesized that participants may have been physically active upon arriving in the room; then signing the consent form, learning about the sensors and devices they would be wearing, may have caused some stress. Hence, when we started the study immediately after, some of the residual physiological responses being experienced by the participants may have continued during the baseline rest period of the study.

To test our hypothesis, we discarded the first 6 minutes of the initial rest period and marked it as a "settle-down" period for the participants. We then used only the last 4 minutes of the rest period as our baseline. We computed features from this baseline rest period, and ran Welsh's t-test. Table 4 clearly shows certain features had a statistically significant difference for the Cold Test, as well, suggesting there may be some truth to our hypothesis. One can also see that the significant features for the Math Test were the same as in Table 3, but the Book Test had another feature showing statistical significance – that is, RMSSD.

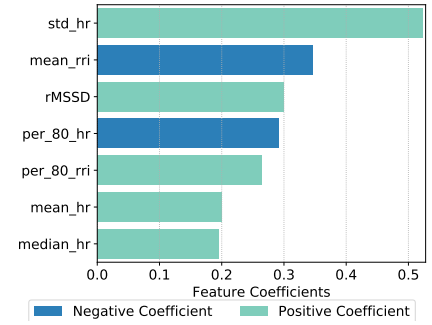
Detecting Stress Events

It is interesting to see that RMSSD (which correlates strongly with High Frequency (HF) bands of heart-rate), is a significant feature for only the Book Test. To understand this, we go back to what RMSSD represents, i.e., the parasympathetic activity, which is the branch of autonomic nervous system in charge of rest functions and recovery.

Here, *recovery* is the key. In the Book Test, we were startling the participants by randomly dropping a heavy book behind them every 30-45 seconds. While the book drop creates an immediate startle response, the participants start recovering from the startled/shocked state immediately after; which is not the case with the Math Test, and the Cold Test, in which the stressors are applied continuously, without giving the participants time for recovery. It is this recovery in the Book Test that is being captured by RMSSD, and likely why it shows a significant difference. We believe this observation is important and may help future researchers working on stress inference and interventions to quantify how well their interventions are working.



(a) Feature importance using Random Forest: The green bars represent the feature importance in the forest, along with their inter-tree variability.



(b) Feature importance using Linear SVM: The green bars represent positive feature coefficient, whereas the blue bars represent negative coefficients.

Figure 2: Feature Importance representation using Random Forest and Linear SVM, only with heart-rate features, sorted from highest to lowest. For the sake of space, we only show the top seven features.

Having determined that the features computed from heart-rate data (as measured by a readily available, commercial, off-the-shelf, heart-rate monitor (the Polar H7)) showed significant differences between rest and stress-induced periods, we next used these features (mentioned in Table 2) to build machine-learning models designed to infer whether the person is *stressed* or *not stressed*. Further, during a stressful period, we look at the feasibility of differentiating among the three types of stressors, i.e., Math, Book, and Cold tests.

Inferring 'stressed' vs. 'not stressed'. We computed features on each one-minute window, and then labeled the window as either 1 (stressed) or 0 (not stressed), based on whether the participant was undergoing a stress induction task during that minute.

In the past, researchers have used several machine-learning algorithms for stress detection; two are widely used and have also been shown to consistently perform better in comparison to others: Support Vector Machines (SVM) and Random Forests (RF) [4, 9, 11, 16, 18]. We used both of these popular machine-learning algorithms in our work, compare their performance, and evaluate how the performance metrics change with different combinations of outlier handling and normalization methods. All the evaluation results reported are using Leave One Subject Out (LOSO) cross-validation.

For each algorithm (SVM and RF), we output a probability that the instance belonged to the *stressed* class. We then threshold the result: if the probability was greater than the threshold, the instance was classified as positive (1), i.e., *stressed*, else it was classified as negative (0), i.e., *not stressed*. This approach allowed us to adjust the threshold to achieve the highest predictive power; in the future, we may consider using the probability to infer the level of stress the participant is experiencing instead of a binary classification.

While we did the training and evaluation for each of the six combinations of outlier handling and normalization methods, we observed that *outlier_minmax* and *outlier_zscore* consistently performed worst (on all three metrics – precision, recall and F1 score) across all six combinations (which was expected, since we *did not* handle outliers in these two combinations, and leaving them as-is in the data could have introduced a bias). Hence, we do not report results from those two combinations, and show comparisons among the other four options.

We began by considering the whole 10 minutes of the baseline rest period as *not stress*, and each of the three 4-minute stress-induction periods as *stress* (we ignored the resting periods between two stress induction tasks, to allow the participants' physiology to return to baseline). These cross-validation results are shown in Table 5. We then considered only the last 4 minutes of the baseline resting period as *not stress*, ignoring the first 6 minutes. The cross-validation results are shown in Table 6.

On comparing the values reported in Table 5 and Table 6, we observe the inference results resonate with the findings in the previous section, i.e., ignoring the first 6 minutes of the initial rest period

Features	Book Test & Math Test		Cold Test & Math Test		Book Test & Cold Test	
	t-stat	*p-value*	*t-stat*	*p-value*	*t-stat*	*p-value*
mean HR	**−18.160**	**<0.001**	**−8.760**	**<0.001**	**−6.080**	**<0.001**
standard deviation HR	0.300	0.762	0.690	0.491	−0.430	0.671
median HR	**−18.260**	**<0.001**	**−8.760**	**<0.001**	**−6.010**	**<0.001**
20th percentile HR	**−16.850**	**<0.001**	**−8.320**	**<0.001**	**−5.720**	**<0.001**
80th percentile HR	**−16.720**	**<0.001**	**−7.990**	**<0.001**	**−5.750**	**<0.001**
mean R-R	**18.510**	**<0.001**	**7.970**	**<0.001**	**6.350**	**<0.001**
standard deviation R-R	**2.900**	**0.004**	1.150	0.250	1.490	0.137
median R-R	**18.110**	**<0.001**	**8.120**	**<0.001**	**6.090**	**<0.001**
max R-R	**10.810**	**<0.001**	**4.580**	**<0.001**	**4.860**	**<0.001**
min R-R	**11.150**	**<0.001**	**5.050**	**<0.001**	**4.120**	**<0.001**
20th percentile R-R	**17.290**	**<0.001**	**8.170**	**<0.001**	**5.940**	**<0.001**
80th percentile R-R	**16.200**	**<0.001**	**7.180**	**<0.001**	**5.540**	**<0.001**
RMSSD	**2.770**	**0.006**	1.470	0.145	1.200	0.231

Table 7: Significance Test between different stress induced minutes. Significant scores ($p < 0.05$) are shown in bold.

	SVM			Random Forest		
Metrics	*'Math'*	*'Book'*	*'Cold'*	*'Math'*	*'Book'*	*'Cold'*
Precision	0.79	0.72	0.94	0.78	0.72	0.63
Recall	0.85	0.76	0.34	0.83	0.77	0.51
F1 Score	0.82	0.74	0.50	0.80	0.74	0.56

Table 8: LOSO Cross-validation results for a multi-class classification amongst stress induced periods, with the *trim_zscore* dataset, using Linear SVM and Random Forest classifiers.

led to better results. This result strengthens our initial hypothesis about residual stress in the initial minutes of the resting baseline.

In Table 6, we observe that the best result was achieved by SVM on the *trim_zscore* combination, i.e., trim outliers, then z-score normalization. It is interesting to note that while Random Forest produced a consistent F1 score of approximately 0.73 (with varying precision and recall) across the different datasets, SVM showed a wide variation of F1 scores: from 0.66 to 0.81. Note also that *trim_zscore* obtained a recall and F1 score slightly better than what was obtained using the ECG-only data in cStress, one of the leading methods in prior stress-detection research [11]. Our result suggests that it is possible to detect stress using a commodity heart-rate sensor, at least in the lab setting.

To further understand the role of different features in the model performance, we present a ranking of the features (in Figure 2) based on the feature importance scores obtained from the Random Forests classifier, and a Linear SVM classifier (since RBF Kernel SVM does not provide a mean to rank feature importance). The features are shown from the highest rank to the lowest.

Differentiating types of stressor. Now that we have demonstrated that it is possible to train a classifier to detect stress, we next seek to determine whether it is possible to distinguish between the different stress inducing tasks. If so, it may eventually be possible to provide meaningful interventions according to the stressor.

We begin by determining which features might best differentiate stressors. We show the results of Welsh's t-test for each feature for each pair of stressors in Table 7. We observe statistically significant differences among the stressors, for many of the features, implying that the different stressors may lead to different physiological responses from the participants.

Given these promising results, we next trained models that seek to classify the type of stressor experienced. Specifically, when a particular window is known to be *stressful*, we trained models that aim to classify the window based on which stressor was experienced during that window. We thus annotated each stress-induction period with a different label: Math Test as 1, Book Test as 2 and the Cold Test as 3. This task was now a three-class classification, and we trained Linear SVM and Random Forest models for a LOSO cross-validation. Table 8 shows the results. From the table we observe that while we obtained high F1 scores for inferring the Math Test and Book Test, that was not the case for the Cold Test. Also, while SVM and Random Forest both produced similar prediction metrics (precision, recall, and F1 score), for Math and Book Tests, they produced widely varying results for the Cold Test: SVM leads to high precision with low recall, whereas Random Forest does not show such a large difference between precision and recall. We need to look further into the modelling of different kinds of stressful periods (beyond the Math Test, Book Test and Cold Test discussed here) to understand this difference, which we leave to future work. Out of curiosity, we considered a two-class classification between Math Test and Book Test, and ignore the Cold Test completely from

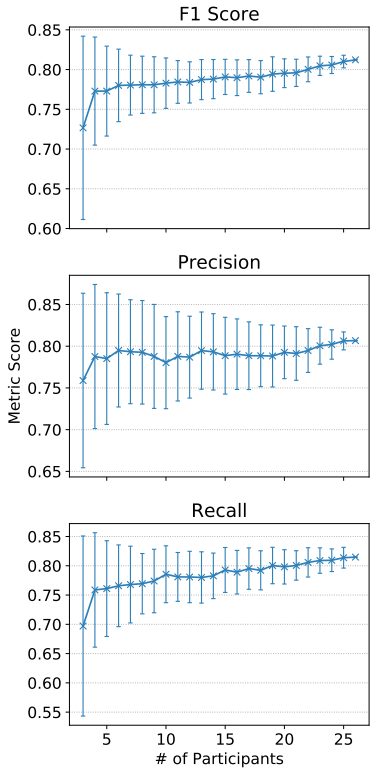


Figure 3: Performance of LOSO cross-validation for detecting stress periods using n number of participants, where $n \in [3, 26]$. The points represents the mean of 200 randomly selected combinations of participants, and the error bars represent standard deviation across different combinations.

[1]For $n = 25$ and $n = 26$, we had 26 and 1 combinations respectively.

[2]The results reported are using the *trim_zscore* combination of the Lab HR data.

the evaluation (from both training and testing). We observed that the F1 score improved significantly for both the classes, with values greater than 0.90 for both.

DISCUSSION AND CONCLUSION

While we have affirmed the possibility of using cheap commodity devices to detect stress, several additional issues need to be explored. We discuss some of these issues in this section.

Order of stress-induction tasks in the lab: Although none of our participants knew what stress-induction tasks would occur during the lab session, the sequence of tasks was the same for all participants, i.e., the Math test, followed by the Book drop test, and finally the Cold water test. While the decision to follow the same sequence of tasks is consistent with previous work [11, 16], some researchers claim that randomizing the order is required to avoid a carry-over effect of previous tasks. Further exploration is required to observe whether and how the order affects the results.

Scalability of the stress detection model: To observe the sensitivity of our model to the number of users used for training the model, we modelled performance for different subsets of users. Considering n to be the number of users in the model, where $n \in [3, 26]$, all the possible combinations of users for each n is ${}^{26}C_n$. For each n, we randomly selected 200 combinations[1] out of the ${}^{26}C_n$ possible combinations and ran a LOSO cross-validation for each combination. We show the *mean* and *standard deviation* of F1 score, Precision and Recall[2] in Figure 3. It is interesting to observe that the metrics, the F1 score for example, varies only slightly after 15–16 participants, with substantial reduction in the standard deviation. While we realize that the convergence at $n = 26$ is because there is only one combination, it is re-assuring to see that the model performance does not vary substantially with different participant combinations. This is, of course, just for a college-student population; we plan to evaluate this in greater detail for a broader population in lab and field situations in future work.

Need for further multi-scale deployment and evaluation: In our work, we looked at the performance of the Polar H7 on 26 participants, in the lab. We recently collected three days of *free-living* data, and hope to report those results in the future; we believe further research with more participants and for longer durations of time is required. In previous work, the cost and availability of custom or clinical-grade commercial devices have limited the reproducibility of studies and large-scale deployments of such devices. We believe the use of cheap commodity devices will help overcome these shortcomings. To this end, we plan on releasing an open-source smartphone app that can collect data from Polar H7 HRM and apply the models described in this paper to infer stress and not-stressed.

ACKNOWLEDGEMENT

This research results from a research program at the Institute for Security, Technology, and Society at Dartmouth College, supported by the National Science Foundation under award numbers CNS-1314281, CNS-1314342, CNS-1619970, and CNS-1619950. The views and conclusions contained in this

document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the sponsors.

REFERENCES

[1] Andrew Baum. 1990. Stress, intrusive imagery, and chronic distress. *Health Psychology* 9, 6 (1990), 653.

[2] George E. Billman. 2013. The LF/HF ratio does not accurately measure cardiac sympatho-vagal balance. (2013), 26 pages. https://doi.org/10.3389/fphys.2013.00026

[3] Biopac Systems Inc. 2016. Biopac MP150. (2016). https://www.biopac.com/wp-content/uploads/MP150-Systems.pdf

[4] Begum Egilmez, Emirhan Poyraz, Wenting Zhou, Gokhan Memik, Peter Dinda, and Nabil Alshurafa. 2017. UStress: Understanding college student subjective stress using wrist-based passive sensing. In *IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE, 673–678. https://doi.org/10.1109/PERCOMW.2017.7917644

[5] Michael R. Esco and Andrew A. Flatt. 2014. Ultra-short-term heart rate variability indexes at rest and post-exercise in athletes: Evaluating the agreement with accepted recommendations. *Journal of Sports Science and Medicine* 13, 3 (Sep 2014), 535–541. http://www.ncbi.nlm.nih.gov/pubmed/25177179

[6] S.M. Fox and W.L. Haskell. 1970. The exercise stress test: needs for standardization. *Cardiology: Current Topics and Progress, New York Academic Press* (1970), 149–154.

[7] Maurizio Garbarino, Matteo Lai, Simone Tognetti, Rosalind Picard, and Daniel Bender. 2014. Empatica E3 - A wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition. In *Proceedings of the International Conference on Wireless Mobile Communication and Healthcare*. ICST. https://doi.org/10.4108/icst.mobihealth.2014.257418

[8] Martin Gjoreski, Hristijan Gjoreski, Mitja Luštrek, and Matjaž Gams. 2016. Continuous Stress Detection Using a Wrist Device: In Laboratory and Real Life. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp'16 Adjunct)*. ACM Press, 1185–1193. https://doi.org/10.1145/2968219.2968306

[9] Martin Gjoreski, Mitja Luštrek, Matjaž Gams, and Hristijan Gjoreski. 2017. Monitoring stress with a wrist device using context. *Journal of Biomedical Informatics* 73 (Sep 2017), 159–170. https://doi.org/10.1016/j.jbi.2017.08.006

[10] Jennifer A. Healey and Rosalind W. Picard. 2005. Detecting stress during real-world driving tasks using physiological sensors. *Intelligent Transportation Systems, IEEE Transactions on* 6, 2 (2005), 156–166.

[11] Karen Hovsepian, Mustafa Al'Absi, Emre Ertin, Thomas Kamarck, Motohiro Nakajima, and Santosh Kumar. 2015. cStress: Towards a Gold Standard for Continuous Stress Assessment in the Mobile Environment. *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)* (2015), 493–504. https://doi.org/10.1145/2750858.2807526

[12] Martin Kusserow, Oliver Amft, and Gerhard Troster. 2013. Monitoring Stress Arousal in the Wild. *IEEE Pervasive Computing* 12, 2 (April 2013), 28–37. https://doi.org/10.1109/MPRV.2012.56

[13] Sylvain Laborde, Emma Mosley, and Julian F Thayer. 2017. Heart rate variability and cardiac vagal tone in psychophysiological research - Recommendations for experiment planning, data analysis, and data reporting. (2017), 213 pages. https://doi.org/10.3389/fpsyg.2017.00213

[14] Amir Muaremi, Agon Bexheti, Franz Gravenhorst, Bert Arnrich, and Gerhard Troster. 2014. Monitoring the impact of stress on the sleep patterns of pilgrims using wearable sensors. In *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE, 185–188. https://doi.org/10.1109/BHI.2014.6864335

[15] NIMH. 2016. Fact Sheet on Stress. (2016). https://www.nimh.nih.gov/health/publications/stress/index.shtml.

[16] K. Plarre, A. Raij, S. M. Hossain, A. A. Ali, M. Nakajima, M. Al'absi, E. Ertin, T. Kamarck, S. Kumar, M. Scott, D. Siewiorek, A. Smailagic, and L. E. Wittmers. 2011. Continuous inference of psychological stress from sensory measurements collected in

the natural environment. In *Proceedings of the IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE, 97–108. http://ieeexplore.ieee.org/xpls/abs

[17] Polar. 2017. Polar H7. (2017). https:llwww.polar.com/us-en/products/accessories/H7

[18] Hillol Sarker, Inbal Nahum-Shani, Mustafa Al'Absi, Santosh Kumar, Matthew Tyburski, Md M. Rahman, Karen Hovsepian, Moushumi Sharmin, David H. Epstein, Kenzie L. Preston, C. Debra Furr-Holden, and Adam Milam. 2016. Finding Significant Stress Episodes in a Discontinuous Time Series of Rapidly Varying Mobile Sensor Data. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI)*. ACM Press, 4489–4501. https://doi.org/10.1145/2858036.2858218

[19] Fred Shaffer and J P Ginsberg. 2017. An Overview of Heart Rate Variability Metrics and Norms. *Frontiers in Public Health* 5 (2017), 258. https://doi.org/10.3389/fpubh.2017.00258

[20] H. Tanaka, K. D. Monahan, and D. R. Seals. 2001. Age-predicted maximal heart rate revisited. *Journal of the American College of Cardiology* 37, 1 (Jan 2001), 153–6. http://www.ncbi.nlm.nih.gov/pubmed/11153730

[21] M. Wu. 2006. *Trimmed and Winsorized Estimators*. Ph.D. Thesis. Michigan State University.

[22] Zephyr. 2018. Zephyr BioHarness 3. (2018). https://www.zephyranywhere.com/media/download/bioharness3-user-manual.pdf