Continuous Detection of Physiological Stress with Commodity Hardware

VARUN MISHRA, GUNNAR POPE, and SARAH LORD, Dartmouth College
STEPHANIE LEWIA, University of Massachusetts Amherst
BYRON LOWENS and KELLY CAINE, Clemson University
SOUGATA SEN, RYAN HALTER, and DAVID KOTZ, Dartmouth College

Timely detection of an individual's stress level has the potential to improve stress management, thereby reducing the risk of adverse health consequences that may arise due to 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 to measure physiological signals; they are often bulky, custom made, and expensive, hence limiting their adoption by researchers and the general public. In this article, we explore the viability of commercially available off-the-shelf sensors for stress monitoring. The idea is to be able to use cheap, nonclinical sensors to capture physiological signals and make inferences about the wearer's stress level based on that data. We describe a system involving a popular off-the-shelf heart rate monitor, the Polar H7; we evaluated our system with 26 participants in both a controlled lab setting with three well-validated stress-inducing stimuli and in free-living field conditions. Our analysis shows that using the off-the-shelf sensor alone, we were able to detect stressful events with an *F*1-score of up to 0.87 in the lab and 0.66 in the field, on par with clinical-grade sensors.

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

Additional Key Words and Phrases: 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. 2020. Continuous Detection of Physiological Stress with Commodity Hardware. *ACM Trans. Comput. Healthcare* 1, 2, Article 8 (April 2020), 30 pages.
https://doi.org/10.1145/3361562

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, and CNS-1619970, CNS-1619950, and by the NIH National Institute of Drug Abuse under award number NIH/NIDA P30DA029926. 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.

Authors' addresses: V. Mishra, S. Sen, and D. Kotz, Department of Computer Science, 6211 Sudikoff Lab, Hanover, NH 03755; emails: varun@cs.dartmouth.edu, {Sougata.Sen, David.F.Kotz}@dartmouth.edu; G. Pope, Thayer School of Engineering, 14 Engineering Dr, Hanover, NH 03755; email: Gunnar.C.Pope.TH@dartmouth.edu; S. Lord, CTBH, 46 Centerra Parkway, EverGreen Center, Suite 315, Lebanon, NH 03766; email: Sarah.E.Lord@dartmouth.edu; S. Lewia, College of Education, 813 N Pleasant St, Amherst, MA 01003; email: slewia@umass.edu; B. Lowens and K. Caine, School of Computing, McAdams Hall, Clemson, SC 29634; emails: blowens@g.clemson.edu, caine@clemson.edu; R. Halter, Thayer School of Engineering, 14 Engineering Dr, Hanover, NH 03755; email: Ryan.J.Halter@dartmouth.edu.
Author's current address: S. Sen, 2145 Sheridan Rd, Department of EECS, Northwestern University, Evanston, IL 60201; email: Sougata.Sen@northwestern.edu.

2637-8051/2020/04-ART8 $15.00
https://doi.org/10.1145/3361562

1 INTRODUCTION

Stress is defined as the brain's response to any demand or change in the external environment [41] and has the potential to actuate changes within an individual's lifestyle. The word *stress* might connote a negative impression, but that is not always the case. Among the different types of stress, *acute stress* is the type of stress that brings excitement, thrill, and the feeling of adrenaline rush into our lives. This is the type of stress that occurs over a very short period of time and can lead to positive outcomes. One example of acute stress may be familiar to readers who are authors: the stress induced by a paper submission deadline often results in the authors being motivated to complete the last few writing or editing tasks that will complete the paper.

However, when an individual experiences sustained stress, it may lead to emotional concerns (e.g., anger, anxiety, or acute periods of depression) and physical distress (e.g., headaches, digestive problems, and even diabetes) [1, 14, 36, 45]. Moreover, if interventions to mitigate this sustained stress are not provided, it could further lead to *chronic stress*, which is severely detrimental to both physical and mental health [5]. Methods for continuous monitoring of an individual's stress level can serve as a foundation for understanding the relationship between stress and behavior, or stress and context. This foundation would scaffold the development of novel interventions that could help individuals recognize and manage stress, or negative behaviors triggered by stress.

Existing methods commonly used by behavioral psychologists to quantify and monitor stress levels, such as the Perceived Stress Scale (PSS) [16], have two limitations: (1) they rely on self-report data, and (2) they are windows into moments in time rather than continuous monitors. Moreover, these methods require respondents to stop their ongoing activity to fill in the questionnaire. These limitations, although acceptable for retrospective studies of stress, make prospective studies and real-time interventions impossible. For real-time interventions, we need to be able to continuously measure and monitor an individual's stress level. One approach to enable this sort of real-time measurement and feedback is through the use of wearable sensors.

With recent advancements in sensor and wearable technologies, it is now possible to continuously collect and stream physiological signals for near-real-time analysis. Indeed, researchers are beginning to make progress on continuous and passive measurement of stress, both in the laboratory and in free-living settings [25, 28, 29, 32, 33, 38, 42, 48]. Although 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. These limitations prevent large-scale adoption of such sensors by (1) researchers who want to observe participant stress in real or near-real time; (2) researchers who want to study interventions and their effect on other behaviors such as anxiety, smoking cessation or drug abuse; and (3) consumers who want to monitor their stress level beyond the clinical setting, in free-living conditions. Toward making accessible and affordable wearable sensors for stress monitoring possible, in this work we aim to answer the following question: *Can we use a commodity device to accurately detect stress?*

While answering this question, we make the following key contributions:

- We demonstrate the feasibility of using *just* a commodity, off-the-shelf heart rate monitoring device (the Polar H7 [43]) to measure stress in controlled (lab) and free-living conditions. Ours is the first work in this direction and is a big step forward from using a clinical-grade sensor or a specialized custom sensor for monitoring physiological stress [15, 26, 32, 39, 42, 46, 47].
- We compare a variety of data processing methods and their effect on the accuracy of stress inference using a commodity sensor. We demonstrate that some of the typical preprocessing steps used in prior work do not perform equally well for commodity devices.
- We make recommendations about the data processing pipeline for the task of stress detection. Although our aim was to test applicability for commodity sensors, we show our pipeline also applies to custom-built sensors (a galvanic skin response (GSR) sensor) as well. We believe that the recommendations made can

be generalized and bring uniformity to the task of stress detection. It is our hope that our work can serve as a guideline to future research on stress detection.

- We propose a novel two-layer method for detecting stress, which can account for a participant's previous stress level while determining the current stress level. We show that using the two-layer approach leads to a notable improvement in stress detection performance. Using only the data from Polar H7 (heart rate and R-R interval), we saw an $F1$-score of 0.88 and 0.66 for detecting stress in the lab and free-living conditions, respectively.

These contributions give us confidence about the usability of commodity heart rate monitors (in this case, the Polar H7) in both lab and field testing conditions, either by itself or in conjunction with other sensors/devices. Although more analyses with a larger, diverse cohort is required, we believe that our work is a strong step toward eliminating researchers' dependence on custom or expensive clinical-grade ECG monitors for stress measurement, and enables study of other mental and behavioral health outcomes.

2 RELATED WORK

Improvements in sensors and sensing capabilities over the years have led to a spectrum of prior work in stress detection and assessment. There are multiple methods that have been used for "contactless" stress measurement, such as using the user's voice [35], or using accelerometer-based contextual modeling [24], or phone usage data like Bluetooth and Call/SMS logs [10]; however, we focus on related works using wearable devices for physiology-based stress measurement. Although contactless approaches have some advantages, they also have several limitations, such as lack of continuous assessment, dependency on personalized models, or the need for extensive training across various situations.

Prior research has attempted to detect stress in a variety of situations. These situations can be broadly classified as (1) stress induced in a lab, where researchers ask the participants to undergo some well-validated stress-inducing tasks [12, 18, 26, 42, 48]; (2) constrained real-life situations, where the researchers monitored the user's stress level in a particular situation, such as in a call center [30], while driving [29], or while sleeping [40]; and (3) in unconstrained free-living situations [26, 32, 42, 47].

To measure stress in the aforementioned scenarios, researchers have used a combination of signal processing, statistical analysis, and machine-learning models on a variety of physiological sensors, such as the respiration sensor (RIP) [26, 29, 32, 42, 48], the electrocardiography (ECG) sensor [26, 32, 42], the GSR[1] sensor [13, 18, 26, 29, 50], the blood volume pulse (BVP) sensor [26], or the electromyogram (EMG) sensor [29].

In several of these works, the researchers developed their own custom-fitted sensing system [13, 29, 32, 42, 50]. 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, yet they also have some major limitations, such as 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, however, a few works that have used a *commercially* available sensor, either by itself [26, 40] or in conjunction with a custom sensor [18] or a smartphone [47].

Muaremi et al. [40] used a combination of the Zephyr BioHarness 3.0 [57] and an Empatica E3 [23] for monitoring stress while sleeping. Gjoreski et al. [26] used both the Empatica E3 and E4 [19] to detect stress in a lab and an unconstrained field (free-living) setting. Sano and Picard used the Affectiva Q Sensor along with smartphone usage data to predict the PSS scores at the end of the experiment [47]. In all of these works, the sensors they used are marketed as "highquality" or "clinicalquality" physiological sensors and hence are too expensive[2] to be

[1] In the research community, GSR is also referred to as electro-dermal activity (EDA). In this work, we use the term *GSR*.

[2] Zephyr BioHarness [58] sells on Amazon for more than $650; Empatica E3 has been replaced with E4, which sells for more than $1,600 and has a 2- to 6-week shipping time [19], and the Affectiva Q Sensor has been discontinued, but according to some blog posts, it used to cost around $2,000 [17].

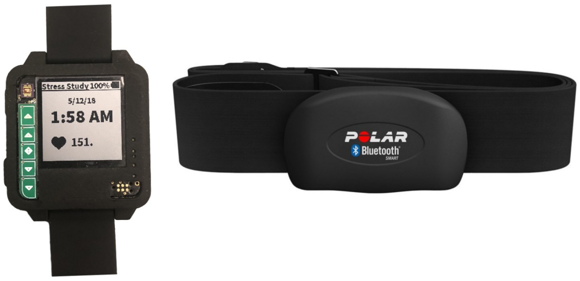


Fig. 1. The devices used in our study: the Amulet wrist device (left) and the Polar H7 chest sensor (right).

considered *commodity* devices. The high cost of these sensors limit large-scale deployments of these devices in studies of stress detection and other mental and behavioral health outcomes. In contrast, we use a *commodity* device, the Polar H7 heart rate monitor [43].[3]

Although the Polar H7 has also been used by Egilmez et al. [18] in UStress, they used it to just get the heart rate values (beats per minute) to act as a supplement to their custom GSR sensor for stress prediction in the lab setting. The authors then compared the differences in prediction results by using heart rate information obtained from a chest-strap sensor (Polar H7) and a smartwatch (from LG). Our work, however, gives insights into the feasibility of using the heart rate and R-R interval data from just a commodity sensor (Polar H7) for being able to detect and predict stress both in a controlled lab environment and an unconstrained field scenario.

We summarize all of the previous work mentioned in this section in Table 1. We report the type of environment/situation(s) where the study was conducted, the type of data collected in those studies, the types of sensors/devices used, the number of participants, and the results obtained by the authors.

3 DATA COLLECTION

We conducted a study, comprising lab and field components, with $n = 27$ participants (15 females, 12 males; 13 undergraduate and 14 graduate students), with a mean age of 23 ± 3.24 years. The study was approved by our Institutional Review Board. All participants completed both the lab and field components and were compensated with $50 for their time.

In what follows, we describe the devices used, the lab and field procedure, and the data collected.

3.1 Wearable Devices

During the course of the study, all participants wore a commercially available, off-the-shelf heart rate monitor (Polar H7 [43]), which is a chest-worn device capable of collecting both the heart rate value (in beats per minute (bpm)) and the R-R interval[4] values (in milliseconds).

In addition to the heart rate monitor, for the field study, the participants wore the Amulet wrist device [31] to collect activity data and trigger ecological momentary assessment (EMA) prompts. The Amulet also served as the data hub to collect the data from the heart rate monitor using Bluetooth Low Energy (BLE). The devices used in the study are shown in Figure 1.

3.1.1 The Heart Rate Monitor. For this study, we wanted to use a commodity heart rate monitor that supported BLE. We chose to use a chest-worn heart rate monitor because the accuracy of chest-strap heart rate monitors

[3] The Polar H7 is available on Amazon for just under $60, and has been recently updated with the newer Polar H10 which is available on Amazon for $70.

[4] R-R interval represents the time interval between two consecutive R peaks in the QRS complex of an ECG wave. R-R interval is a measure of interbeat variability, also known as heart rate variability (HRV), and has been shown to be a marker for stress and health [53].

Table 1. Summary of Related Work

	Situation	Subjects (#)	Types of Data Used	Devices Used for Data Collection	Prediction Metrics Reported
Choi et al. [13]	Lab	10	HRV, RIP, GSR, and EMG	Custom chest–strapped sensor suite	Binary classification between stressed and not stressed with 81% accuracy
Healey and Picard [29]	Driving tasks	9	EKG, EMG, respiration, GSR	Custom sensors	Classification between low, medium, or high stress with 97% accuracy
Hernandez et al. [30]	Call center	9	GSR	Affectiva Q Sensor	Personalized model: 78.03% accuracy Generalized model: 73.41% accuracy
Muaremi et al. [40]	Sleeping	10	ECG, respiration, body temperature, GSR, upper body posture	Empatica E3, Zephyr BioHarness 3.0	Classification between low, moderate, or high stress with 73% accuracy
Egilmez et al. [18]	Lab	9	Heart rate, GSR, gyroscope	Custom GSR sensor with LG smartwatch	Binary classification: *F*1-score of 0.888
Sano and Picard [47]	Field	18	GSR and smartphone usage	Affectiva Q Sensor, and smartphones	Binary classification, 10-fold cross validation: 75%
Plarre et al. [42]	Lab, field	21	ECG and RIP	Custom sensor suite, AutoSense	Lab: Binary classification of stress with 90.17% accuracy, Field: High correlation ($r = 0.71$) with self-reports
Hovsepian et al. [32]	Lab, field	Lab train data: 21 participants Lab test data: 26 participants Field test data: 20 participants	ECG and RIP	Custom sensor suite, AutoSense	Binary classification of stress: Lab train LOSO CV *F*1-score: 0.81 Lab test *F*1-score: 0.9 Field self-report prediction *F*1-score: 0.72
Sarkar et al. [48]	Field	38	ECG and RIP	Custom sensor suite, AutoSense	Using models generated with cStress, field self-report prediction *F*1-score: 0.717
Sun et al. [50]	Lab	20	ECG and GSR	Custom chest and wrist-based sensor suite	Binary classification of stress, accuracy by 10-fold cross validation: 92.4%; accuracy for cross-subject classification: 80.9%
Gjoreski et al. [25]	Lab, field	Lab: 21 Field: 5	BVP, GSR, HRV, skin temperature, accelerometer	Empatica E3 and E4	Lab: Classification between no stress, low stress, and high stress achieved 72% LOSO accuracy Field: Binary classification for detecting stress with *F*1-score of 0.81

Table 2. Pearson's Correlation of Features Computed by Zephyr HXM and Polar H7, with the Lab Benchmark—Biopac

	Zephyr HXM		Polar H7	
Features	*r-Coefficient*	*p-Value*	*r-Coefficient*	*p-Value*
Mean of heart rate (HR)	0.778	<0.001	0.917	<0.001
Standard deviation HR	0.361	0.077	0.809	<0.001
Median HR	0.759	<0.001	0.794	<0.001
20th percentile HR	0.831	<0.001	0.838	<0.001
80th percentile HR	0.704	<0.001	0.951	<0.001
Mean R-R interval	0.768	<0.001	0.999	<0.001
Standard deviation R-R	0.622	0.001	0.951	<0.001
Median R-R	0.757	<0.001	0.988	<0.001
Max of R-R	0.533	0.006	0.663	0.002
Min of R-R	0.595	0.002	0.855	<0.001
20th percentile R-R	0.661	0.000	0.995	<0.001
80th percentile R-R	0.966	<0.001	0.997	<0.001

is better than with wrist-worn optical sensors [3, 4]. We wanted a heart rate monitor that supported BLE so the data could stream to the Amulet. At the time, there were two popular BLE-capable, chest-mounted heart rate monitors available on the market: the Zephyr HXM and the Polar H7. Both are capable of streaming data and follow the standard Heart Rate Profile protocol specifications.[5] Both devices transmit data to the Amulet using BLE at 1 Hz. Each data packet consists of one heart rate value and one or more R-R interval values.

We conducted a preliminary test to compare these heart rate monitors to a popular clinical ECG device—the Biopac MP150 [7]. We first measured participants with the Zephyr and the Biopac, and then with the Polar H7 and the Biopac. We then divided the data collected from each device pair into 30-second windows and computed some basic heart rate and R-R interval features. We used a Pearson correlation to compare the feature values between the two devices; the results are shown in Table 2. On inspecting the *r*-coefficients from the two comparisons, we observe that the features computed from the Biopac were more strongly correlated with the Polar H7 than with the Zephyr HXM. Given the better performance of the Polar H7, we used it for the study.

3.1.2 The Amulet Wearable Platform. The Amulet is an open source hardware and software platform for writing energy- and memory-efficient sensing applications [9, 31]. The Amulet has several on-board sensors and peripherals, including a three-axis accelerometer, light sensor, ambient air temperature sensor, buttons, capacitive touch slider, micro-SD cards, LEDs, and a low-power display. We used the Amulet to act as a data hub to receive the heart rate data using BLE, to record accelerometer data and to prompt Ecological Momentary Assessment (EMA) questions to the participant. The Amulet stored all of the sensor data and EMA responses on its internal micro-SD card. The rationale behind using the Amulet for *in-the-wild* data collection and storage instead of a smartphone was that a wearable would always be on the body of the participant, thus reducing the chances of data loss when the phone was not in range of the person. In addition, the Amulet had additional physical buttons that we were able to map for specific tasks, as described in Section 3.4. Finally, a wearable like Amulet can collect data about the participant's stress and physical activity even when the smartphone is on the table, in another room, or being used by someone else.

[5] https://www.bluetooth.org/docman/handlers/downloaddoc.ashx?doc_id=239865.

3.2 Lab Study

The purpose of the lab study was twofold: first, to determine whether a commodity sensor could be used to detect stress, and second, to establish ground-truth information about the physiological effects of stress, as recorded by the wearables.

We first described the details of the study to the participants, and they consented to participate. Next, participants put on both the heart rate sensor and the Amulet (which was used solely to to collect the data from the heart rate sensor). Once the devices were in place, we began data collection. Each device collected data throughout the lab experiment (about 40 minutes). Participants were asked to not move, remove, or interact with the sensors in any way.

Next, the participant experienced three types of stressors—mental arithmetic, startle response, and cold water—all well-validated stimuli known to induce stress. Specifically, the protocol was as follows:

(1) *Resting baseline*: Participant sat in a resting position for 10 minutes.
(2) *Mental arithmetic task*: Participant counted backward in steps of 7 (4 minutes).
(3) *Rest period*: Participant sat in a resting position for 5 minutes to allow him or her to return to baseline.
(4) *Startle response test*: Participant faced away from the lab staff and closed his or her eyes; staff then dropped a book at several random and unexpected moments, startling the participant (4 minutes).
(5) *Rest period*: 5 minutes, as before.
(6) *Cold water test*: Participant submerged his or 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.

At the end of the initial baseline rest period and after each stressor, we asked the participant to verbally rate his or her stress level on a scale from 1 to 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.

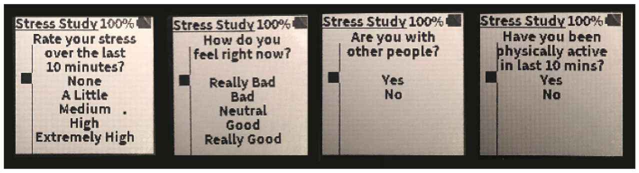
3.3 Field Study

For the field study, participants were asked to wear the sensing system for 3 days (at least 8 hours per day) while carrying out their everyday activities. To ensure that the battery of the devices did not drain before the end of the day, we duty cycled the sensing system to record the physiological data for 1 minute every 3 minutes. In addition to recording the physiological data, the Amulet prompted the participants to answer EMA questions once every 30 minutes. The participants also had the option to proactively report a "stressful" event by clicking on a dedicated "event mark" button. When the participant clicked on this button, the Amulet would record the time as a stressful event, and (in a fraction of such cases) the Amulet would randomly prompt the participant to complete an EMA questionnaire. If an EMA prompt was triggered due to an event mark, the system would not trigger another EMA again in that 30-minute period, to prevent participant overload.

At the end of the field study, participants returned the devices to the lab, completed an exit questionnaire, and were compensated $50.

3.4 Self-Report Data

We asked the participants to self-report their stress levels during both the lab and field segments. In the lab, we asked participants to report their stress level on a scale from 1 to 5 every stress-inducing period (for a total of four self-reports). In the field, the Amulet prompted participants to answer EMA questions every 30 minutes. The participants were instructed to answer every EMA prompt; if any particular prompt was not answered within 5 minutes, it would disappear and was recorded as unanswered.



Stress Study 100%
Rate your stress
over the last
10 minutes?
None
A Little
Medium
High
Extremely High

Stress Study 100%
How do you
feel right now?
Really Bad
Bad
Neutral
Good
Really Good

Stress Study 100%
Are you with
other people?
Yes
No

Stress Study 100%
Have you been
physically active
in last 10 mins?
Yes
No

Fig. 2. EMA prompts shown on the Amulet.

The EMA prompt asked a 5-poing Likert scale question to get an estimate about the participant's stress level: "Rate your stress over the last 10 minutes." Fredriksson-Larsson et al. [22] observed that a single item (5-point Likert scale) stress question has a positive correlation ($r = 0.569, p < 0.001$) with the PSS (a 10-item survey, which is an established method for measuring stress).

In addition to the single-item stress question, each EMA prompt included three additional questions: "How do you feel right now?" (on a scale of 1 to 5), "Are you with other people?" (yes/no), and "Have you been physically active in the last 10 minutes?" (yes/no).

The questions, as they appeared on the Amulet, are shown in Figure 2. The participants used the Amulet's capacitive slider input to scroll between the choices and then click on the bottom-left button to confirm their choice and move on to the next question. Each of the responses were timestamped and used as ground truth for the field condition.

3.5 Data Collected

While we attempted to collect heart rate variability (HRV), EMA, and accelerometer data for 27 participants in the lab and field, there were problems that led to loss of some data. We ran into a problem with corrupt SD cards, which led to partial field data loss for 2 participants. We also lost the complete lab data for 1 participant, due to the same SD card problem. Eventually, we ended up with 26 participants for whom we had heart rate data both in the lab and the field.

The participants were reasonably compliant with responding to the EMA prompts. We received a total of 1,246 valid EMA responses (mean ≈ 46 responses per participant); we explain what it means to be "valid" in the next section. We also received 536 event marks when the participants indicated a stressful situation, which translated to an average of almost 20 events per person.

Figure 3 provides a summary of the data collected; we quantify the amount of lab, field, and EMA data collected from each participant.

4 DATA PROCESSING

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

4.1 Getting Data from Devices

Once the participants returned the devices to us, we extracted the data from those devices. The Amulet logged the heart rate, R-R interval, accelerometer, EMA, and event mark data to files on the built-in micro-SD memory card. The Amulet encrypted these files as they were written to ensure the data was not compromised in case a participant lost the Amulet and/or the micro-SD card. We decrypted the data using a Ruby script that also generates a ".csv" data file for each participant.

4.2 Data Cleaning

We began with preliminary data cleaning to filter out invalid data points. In this step, we were not trying to handle outliers (which may or may not be valid readings) but wanted to remove obviously erroneous data readings. This

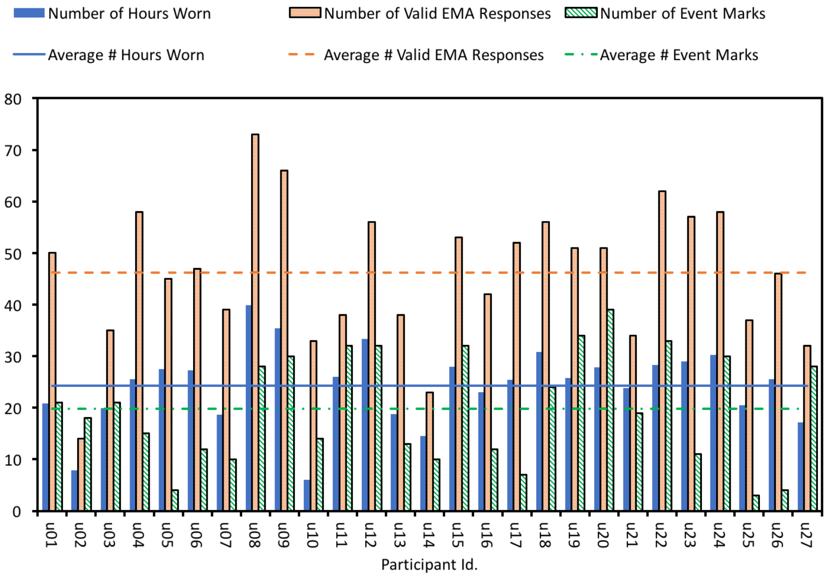


Fig. 3. Summary of the data collected during the field component of the study.

step was important because the sensors used for physiological measurements are not clinical quality and may need a "hot-fix"[6] period. We noticed that 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 predetermined range, it was considered as noise and removed. We dropped both the heart rate value and any R-R interval values received in that second. Based on previous research conducted to find the maximum human heart rate [21, 51], we set our upper bound to 220 bpm. To determine the lower bound, we inspected heart rate data of all 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.

4.3 Self-Report Quality Check

Although the validity and reliability of self-reports have been questioned [2], lack of a better measure has led to self-reports (EMAs) becoming the "gold standard" for the ground truth in many field-based studies [26, 32, 48], not only for stress detection but also for a variety of mental and behavioral health outcomes [54, 55].

Prior researchers have used the time taken to respond as a measure to determine whether the response to an EMA prompt is valid, based on the time taken to answer the EMA [55], or metrics such as Cronbach's alpha to measure the internal consistency in a participant's response to a self-report [32].

In our work, we use a twofold check: first, if the participant selected the default choice presented for each question (1-1-1-1), then we treat that report as invalid, and second, in the self-reports where the default choices were not selected, if the participant completed the report in less than 3 seconds (1 second each for the 5-point Likert scale questions, and 0.5 seconds for the "yes/no" questions), we discard that report. After this filter, we had a total of 1,246 *valid* responses.

Further, we assess the consistency of responses to the self-reports. In the field component, two of the questions in each report were 5-point Likert scale questions: (a) "Rate your stress over the last 10 minutes" (ranging from

[6] Time taken to get stable heart rate readings; we observed that the Polar would sometimes need a few seconds to acquire the heart rate readings.

Table 3. Pearson's Correlation of the First Two EMA Questions Asked During the Field Study, Along with the Individual and Overall Cronbach's Alpha

Participant ID	Correlation Between the EMAs		Cronbach's Alpha
	r-Coefficient	*p-Value*	
u01	**−0.720**	**<0.001**	0.837
u02	**−0.443**	**0.008**	0.614
u03	**−0.408**	**0.001**	0.580
u04	**−0.591**	**<0.001**	0.743
u05	**−0.552**	**<0.001**	0.711
u06	−0.130	0.274	0.230
u07	−0.161	0.196	0.278
u08	**−0.388**	**0.026**	0.559
u09	**−0.590**	**<0.001**	0.742
u10	**−0.808**	**<0.001**	0.894
u11	**−0.802**	**<0.001**	0.890
u12	−0.161	0.464	0.277
u13	**−0.717**	**<0.001**	0.835
u14	−0.134	0.398	0.236
u15	**−0.623**	**<0.001**	0.768
u16	**−0.373**	**0.005**	0.543
u17	**−0.520**	**<0.001**	0.684
u18	**−0.500**	**<0.001**	0.667
u19	**−0.788**	**<0.001**	0.881
u20	−0.165	0.201	0.283
u21	−0.090	0.507	0.164
u22	**−0.533**	**<0.001**	0.695
u23	−0.151	0.372	0.263
u24	**−0.676**	**<0.001**	0.806
u25	**−0.541**	**0.001**	0.702
u26	**−0.788**	**<0.001**	0.881
u27	0.053	0.856	0.101
All participants	**−0.552**	**<0.001**	0.711

"None (1)" to "Extremely High (5))" and (b) "How do you feel right now?" (ranging from "Really Bad (1)" to "Really Good (5))." Assuming that both questions are measuring the same underlying trait (i.e., a participant's stress), we expect to see a negative correlation between the responses to these two questions across all participants. A Pearson correlation between these two items resulted in $r = -0.551$, $p < 0.001$. Although a correlation coefficient makes sense intuitively, we also looked at the Cronbach's alpha measure for the two questions and found the average $\alpha = 0.711$. This result gives us confidence that the reports obtained in the field are reliable overall.

We also looked at the reports for each individual participant and calculated the two measures—Pearson correlation and Cronbach's alpha (in Table 3)—and observed that although most of the participants have high Pearson r and Cronbach's α values, it was not the case for some participants where both measures are extremely low. This result suggests that some participants might not have been diligent in answering the self-reports and/or might have misunderstood the questions. We, however, did not remove the self-reports from these users from our analyses, since the EMA questions do not *actually* measure the same base trait, and participants might be *stressed* but still do not feel *bad* about it and vice versa.

As reported in Section 3.5, we received 536 event marks when the participants just clicked on a dedicated button on the Amulet to mark a stressful event. Although participants found it to be easy to mark an event as stressful, we fear that it might have been too easy. During the exit interview (i.e., when the participants came back to return the devices after the field setting), some of the participants complained about "accidental clicks" on the event mark button, and without an option of undoing the mistake, the Amulet marked the time as a stressful event. We investigated further to determine the extent of the problem. While answering the self-reports, the participants had to choose their stress level on a scale from 1 to 5. For each participant, we calculated the mean score to all reports they answered. If a participant's self-report value was higher than his or her mean, then we labeled that instance as a *stressed* instance (class = 1); otherwise, we labeled it as *not stressed* (class = 0). Now, according to the study design, the participants might randomly receive a prompt to complete a self-report after they click on the event mark button. Of the 536 instances we received, the participants were prompted to complete the self-report 300 times. Of these 300 instances, the participants chose a stress value greater than their individual mean score only 112 times. This suggests that more than 60% of the time, the participants might have clicked on the event mark button by mistake. Without any means to validate if the remaining 236 instances are genuine, we decided not to use the data collected by event marks for training or evaluating our model. We intend to fix this problem in future studies.

4.4 Handling Activity Confounds

Using physiological signals to detect stress has its own drawbacks. The physiological response to mental stress is similar to that exhibited due to physical activity and strain. Hence, it is imperative to be able to distinguish whether an observed physiological arousal was due to mental stress or just physical activity. To this end, we collect accelerometer data from the Amulet along with the physiological readings. We use an activity-detection algorithm developed for the Amulet in a study with 14 undergraduate participants [8].

The activity-detection algorithm uses the accelerometer data and, for every second, infers one of six different activities: lying down, standing, sitting, walking, brisk walking, or running. For every minute in the field data, we determined the dominant activity level—low, medium, or high—based on the activity for each second. If the activity level for a 1-minute window was low (lying down, sitting, or standing), then we included that window in our analyses.

4.5 Feature Computation

We next use the data remaining after the previous steps to compute features to quantify HRV. We split the data into 1-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.

4.5.1 Outliers. While dealing with outliers in data, the common approaches are (1) leave them in the data, (2) reduce the effect the outliers might have, or (3) 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*[7] to reduce the effect of outliers on the

[7] Winsorization is a method of removing the effect of spurious outliers. In winsorization, instead of removing the outliers, they are replaced with certain percentiles. In our case, values greater than the $upper\_bound$ were replaced with the $upper\_bound$ and values less than the $lower\_bound$ were replaced with the $lower\_bound$.

Table 4. Nomenclature for All of the Different Combinations of Outlier Handling and Normalization Methods

	Outliers Present	**Winsorization**	**Trimming**
Minmax Normalization	outlier_minmax	wins_minmax	trim_minmax
***z*-Score Normalization**	outlier_zscore	wins_zscore	trim_zscore

dataset [56]. This approach was also used by some of the previous works, such as cStress [32] and the work by Gjoreski et al. [26]. For the third approach, 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 set the threshold at 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 more than 99% of the data is unaltered. Having defined *outlier*, we establish the upper and lower bounds as follows:

$$lower\_bound = median - 3 \times MAD,$$

$$upper\_bound = median + 3 \times MAD.$$

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 was done individually for each participant.

4.5.2 Normalization. *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, such as participant-specific mean and standard deviation, so the second normalization technique we tried is 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 that 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 4 provides our nomenclature for each of the methods we used.

4.5.3 Feature Computation. We grouped the normalized data into 1-minute windows. Given the short duration of our lab experiments, we wanted to select the shortest possible window size. Esco and Flatt [20] 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. Furthermore, the 1-minute window has been common in physiological monitoring [29, 32, 42].

For the HRV data, we selected only the time-domain features for our work, as shown in Table 5. All of these time-domain features have been shown to be effective in predicting stressful periods by other researchers [32].

Table 5. All of the Features Computed from the Filtered and Normalized HRV Data Segregated by the Base Measures: Heart Rate and R-R Interval

Heart Rate	R-R Interval
Mean, median, max, min, standard deviation, 80th percentile, 20th percentile	Mean, median, standard deviation, max, min, 80th percentile, 20th percentile, root mean square of successive differences (RMSSD)

Unlike earlier work, however, we actively avoid frequency-domain features (e.g., low-frequency (LF) bands, high-frequency (HF) bands, and low:high frequency (LF:HF) ratio) for the following reasons.

The root mean square of successive differences (RMSSD) of successive R-R intervals is associated with short-term changes in the heart and is considered to be a solid measure of vagal tone and parasympathetic activity, similar to HF [34]. Several studies have also shown that RMSSD and HF are highly correlated [49]. Further, unlike HF, RMSSD is easier to compute and is not affected by other confounding factors such as breathing. Hence, we felt that RMSSD would be a good alternative to HF, thereby nullifying the need to compute HF.

Unlike HF, which represents parasympathetic activity, LF is less clear. Although some researchers believe that LF represents sympathetic activity, others suggest that it is a mix of both sympathetic and parasympathetic activities [6]. Furthermore, the rationale behind using the LF:HF ratio is that since HF represents parasympathetic activity, a lower HF will increase the ratio, suggesting more stress; however, since the role of LF is not really clear, looking at the ratio might be misleading as well [6]. In addition, 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 [32]. Hence we decided to leave out LF and LF:HF features from our work, thus not requiring us to calculate any frequency-domain features.

5 EVALUATION

In this section, we evaluate our approach. We begin by determining whether we were able to capture a significance 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.

5.1 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 Welch'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 earlier, we followed three ways of handling outliers and two ways for data normalization, leading to a total of six combinations, as shown in Table 4. Across all 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 6. It is evident that for the math test and the book test, there are several features that showed statistically significant differences. This, however, is not the case for the cold test, where we found no feature showing statistically significant 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 from 1 to 5), during the lab study, after the baseline rest period, and after each

Table 6. Significant Heart Rate–Based Feature Differences from the Initial Rest Period of 10 Minutes

Features	Math Test		Book Test		Cold Test	
	t-Stat	*p-Value*	*t-Stat*	*p-Value*	*t-Stat*	*p-Value*
Mean heart rate (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
Max HR	**−13.261**	**<0.001**	**3.646**	**<0.001**	−1.913	0.059
Min HR	**−10.786**	**<0.001**	**7.709**	**<0.001**	−0.570	0.570
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 interval	**7.020**	**<0.001**	**−5.830**	**<0.001**	−0.770	0.443
Standard deviation R-R interval	0.220	0.830	−0.350	0.726	1.140	0.254
Median R-R interval	**6.870**	**<0.001**	**−6.760**	**<0.001**	−0.380	0.704
Max R-R interval	**6.790**	**<0.001**	**−6.740**	**<0.001**	0.200	0.843
Min R-R interval	**2.650**	**0.009**	0.180	0.858	0.680	0.496
20th percentile R-R interval	**3.630**	**<0.001**	**−3.270**	**0.001**	−1.780	0.076
80th percentile R-R interval	**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

Significant scores ($p < 0.05$) are shown in bold.

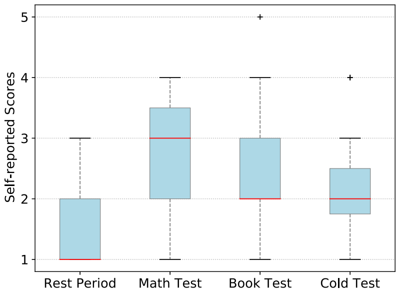


Fig. 4. Participant self-reports after each lab period.

of the stress tests, as shown in Figure 4. 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 Welch's t-test. Table 7 clearly shows that certain features had a statistically significant difference for the cold test as well, suggesting that 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 6, but the book test had another feature showing statistical significance—that is, RMSSD.

Table 7. Significant Heart Rate–Based Feature Differences from the Last 4 Minutes of the Initial Rest Period

Features	Math Test		Book Test		Cold Test	
	t-Stat	*p-Value*	*t-Stat*	*p-Value*	*t-Stat*	*p-Value*
Mean heart rate (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
Max HR	**−14.771**	**<0.001**	**3.194**	**0.002**	**−2.306**	**0.022**
Min HR	**−13.984**	**<0.001**	**8.171**	**<0.001**	−0.757	0.450
20th percentile heart-rate	**−11.640**	**<0.001**	**6.240**	**<0.001**	−1.160	0.249
80th percentile heart-rate	**−12.810**	**<0.001**	**5.030**	**<0.001**	**−2.040**	**0.044**
Mean R-R interval	**13.920**	**<0.001**	**−6.380**	**<0.001**	**2.110**	**0.037**
standard deviation R-R interval	0.350	0.725	**−2.440**	**0.015**	−0.790	0.428
Median R-R interval	**14.150**	**<0.001**	**−6.250**	**<0.001**	1.970	0.052
Max R-R interval	**6.760**	**<0.001**	**−4.960**	**<0.001**	1.130	0.262
Min R-R interval	**8.730**	**<0.001**	−1.760	0.080	**2.460**	**0.015**
20th percentile R-R interval	**13.440**	**<0.001**	**−4.420**	**<0.001**	**2.410**	**0.017**
80th percentile R-R interval	**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

Significant scores ($p < 0.05$) are shown in bold.

It is interesting to see that RMSSD (which correlates strongly with HF bands of heart rate) is a significant feature for only the book test. To understand this, we go back to what RMSSD represents—that is, 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 to 45 seconds. Although 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 that this observation is important and may help future researchers working on stress inference and interventions to quantify how well their interventions are working.

5.2 Evaluation in the Lab Setting

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 5) 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: math, book, and cold tests.

5.2.1 Inferring Stressed Versus Not Stressed. We computed features on each 1-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) [18, 26, 32, 39, 42, 48]. One reason two very different algorithms like SVM and RF are preferred is that both algorithms tend to limit overfitting and reduce the bias and variance of the resulting models. SVMs do so by use of a kernel function and regularization of parameters, whereas RF is an

Table 8. LOSO Cross Validation Results from the Different Datasets Using SVM and RF, and Considering the Entire Rest Baseline of 10 Minutes as *Not Stressed*

Prediction Metrics	*trim_zscore*		*trim_minmax*		*wins_zscore*		*wins_minmax*	
	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
*F*1-score	0.68	0.64	0.56	0.68	0.63	0.63	0.53	0.65

Table 9. 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*

Prediction Metrics	*trim_zscore*		*trim_minmax*		*wins_zscore*		*wins_minmax*	
	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	0.74	0.59	0.67	0.69	0.68	0.59	0.67
*F*1-score	0.81	0.76	0.69	0.73	0.73	0.72	0.66	0.72

ensemble-based classifier, which considers a set of high-variance, low-bias decision trees to create a low-variance and low-bias model. We used both of these popular machine-learning algorithms in our work, compared their performance, and evaluated how the performance metrics change with different combinations of outlier handling and normalization methods.

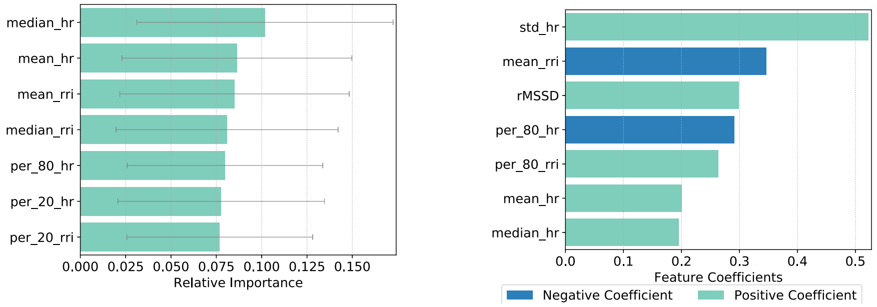
For each algorithm (SVM and RF), we output a probability that the instance belonged to the *stressed* class. We then threshold the result: if the probability was greater than the threshold, the instance was classified as positive (1) (i.e., *stressed*), and otherwise it was classified as negative (0) (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 just a binary classification.

We evaluated each classifier for all the six dataset combinations as mentioned in Table 4 and report three metrics: *precision*, the fraction of those instances labeled "positive" that actually are positive instances; *recall*, the fraction of positive instances labeled correctly as positive; and *F1-score*, the harmonic mean between precision and recall. The *F*1-score is a popular metric in classification problems with one primary class of interest.

For SVM, we used the radial basis function (RBF), which has two hyperparameters: C and γ, the choice of which can significantly affect the results of the SVM algorithm. To choose the best values of the hyperparameters, we performed a grid search and evaluated the performance by leave-one-subject-out (LOSO) cross validation. Basically, for each pair of C and γ, we performed LOSO cross validation for all participants and reported results from the best-performing model. We also tuned the classification threshold. In all cases, we sought to optimize for the *F*1-score by LOSO cross validation.

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 the worst (on all three metrics: precision, recall, and *F*1-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* (ignored the resting periods between two stress induction tasks to allow the participants' physiology to return to the baseline). These cross-validation results are shown in Table 8. 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 9.



(a) Feature importance using RF. The green bars represent the feature impor-tance in the forest, along with their intertree variability.

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

Fig. 5. Feature Importance representation using RF 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.

On comparing the values reported in Table 8 and Table 9, we observe that the inference results resonate with the findings in Section 5.1—that is, ignoring the first 6 minutes of the initial rest period led to better results. This result strengthens our initial hypothesis about residual stress in the initial minutes of the resting baseline.

In Table 9, we observe that the best result was achieved by SVM on the *trim_zscore* combination—that is, trim outliers, then z-score normalization. It is interesting to note that while RF 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.

To further understand the role of different features in the model performance, we present a ranking of the features (in Figure 5) based on the feature importance scores obtained from the RF classifier and a linear SVM classifier (since the RBF kernel SVM does not provide a mean to rank feature importance). The features are shown from the highest rank to the lowest.

5.2.2 Accounting for Prior Stress. In the past, research in the domain of stress detection has focused on computing features in a given window (or sliding window) of time and training machine-learning models to detect stress. This detection could be accomplished either by a direct binary classification or by estimating the probability of stress. For the latter, researchers have used some threshold to classify between binary stress states.

In our work, we followed a similar approach where we used a threshold to classify between the binary stress states. This method assumes that each window of time is independent, which simplifies the building of models for stress detection. By considering each window independently, however, we miss vital information about the previous stress state that could be useful for making an inference about the current window. To this end, we designed a novel two-layer approach for stress detection that accounts for stress in the previous window before making an inference about the current window. The first layer is the estimation of the stress state in the current window, as we have done earlier. The second layer is a Bayesian network model that considers the stress state of the previous window along with the stress state of the current window, detected from the first layer, to infer a final stress state for the current window. Figure 6 illustrates our two-layer approach.

For any given window i, we consider two stress states: S_i, the sensed stress state from the machine-learning model in layer 1, and D_i, the final corrected stress state determined by the Bayesian network model. The Bayesian network model formalizes a recursive relationship between the detected stress state at any given time window (D_i) and the detected stress state at the previous time window (D_{i-1}), and the sensed stress state for that window (S_i).

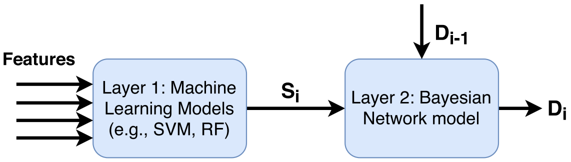

Features
Layer 1: Machine Learning Models (e.g., SVM, RF)
S_i
D_{i-1}
Layer 2: Bayesian Network model
D_i

Fig. 6. Our proposed two-layer approach for stress detection in a given time window i.

Table 10. Conditional Probability Table for the Bayesian Network Model

		D_i	
D_{i-1}	S_i	**0**	**1**
0	0	1	0
0	1	α	$1-\alpha$
1	0	β	$1-\beta$
1	1	0	1

To simplify the parameterization of $p(D_i|D_{i-1}, S_i)$, we make the following assumptions: (1) if the binary stress states at S_i and D_{i-1} are both true, we set the final detected stress state (D_i) as true; (2) if the binary stress states at S_i and D_{i-1} are both false, we set the final detected state to be false. These assumptions are logical and help simplify the model to two parameters, α and β, as outlined in the conditional probability table in Table 10.

Further, the probability from the Bayesian model $p(D_i)$ at any given window i can be marginalized from the joint distribution $p(D_i, D_{i-1}, S_i)$ as

$$p(D_i = 1) = \sum_{k,l=\{0,1\}} p(D_i = 1|D_{i-1} = k, S_i = l) \cdot p(D_{i-1} = k) \cdot p(S_i = l). \tag{1}$$

Considering $p(D_i = 1)$ as y_i, and $p(S_i = 1)$ as x_i, and using the CPT in Table 10, Equation (1) can be simplified as

$$p(D_i = 1) = y_i = (1-\alpha)(1-y_{i-1})x_i + (1-\beta)y_{i-1}(1-x_i) + y_{i-1}x_i. \tag{2}$$

For the first time window ($i = 0$), since we do not have any information about the previous stress state, we set D_0 to be the sensed stress state S_0. Hence, the preceding recurrence in Equation (2) can be initialized as

$$p(D_0 = 1) = y_0 = p(S_0 = 1) = x_0. \tag{3}$$

Now, using Equations (2) and (3), we could calculate the marginal probability of detected stress for any given window. We used this probability of stress to classify between *stressed* and *not-stressed*. We used a grid search to estimate the parameters of α and β, based on the LOSO cross-validation performance. Using the additional layer, we observed that there was a substantial increase in the performance of stress detection. We consider the best-performing model from earlier as the model in layer 1 and show the classification performance to highlight the improvement in Table 11. Using the proposed two-layer model with the Bayesian network model, we observe that the $F1$-score improved by 6 percentage points to 0.87 as compared to the traditional one-layer approach that uses just a machine-learning model. A major benefit of the layered approach is that the Bayesian network model can be applied in conjunction with any classifier in the first layer. In our work, we report results from the base classifiers first, followed by the best-performing classifier in conjunction with the Bayesian network model, to show the improvement in performance.

To put the results obtained into perspective, we compare to results obtained in previous studies in similar situations. We do understand that a direct comparison might not be perfectly appropriate because of different study logistics and demographics. We compare our results using the Polar H7 with the results obtained in cStress [32] (which had a comparable lab protocol) in Figure 7. We compare results reported by using all features (both ECG and respiration features) and results using only the ECG features. It is encouraging to see from the comparison that the $F1$-score obtained by features from a commodity device like Polar H7 is similar to or better than what was reported using high-quality custom sensors in cStress, one of the leading methods in prior stress detection

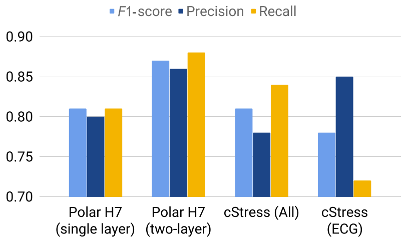


Fig. 7. Comparing stress detection results for Polar H7 with cStress: all features (ECG and respiration) and ECG-only features.

Table 11. Performance Comparison with *trim_zscore*: Using Just a Layer 1 Approach as Before (SVM) Compared to the New Proposed Two-Layer Model

	SVM Only	*Proposed (two-layer)*
Precision	0.80	**0.86**
Recall	0.81	**0.88**
*F*1-score	0.81	**0.87**

Table 12. Significance Test Between Different Stress-Induced Minutes

Features	Book Test & Math Test		Cold Test & Math Test		Book Test & Cold Test		One-Way ANOVA	
	t-Stat	*p-Value*	*t-Stat*	*p-Value*	*t-Stat*	*p-Value*	*F-Stat*	*p-Value*
Mean heart rate (HR)	**−18.160**	**<0.001**	**−8.760**	**<0.001**	**−6.080**	**<0.001**	**153.480**	**<0.001**
Standard deviation HR	0.300	0.762	0.690	0.491	−0.430	0.671	0.250	0.778
Median HR	**−18.260**	**<0.001**	**−8.760**	**<0.001**	**−6.010**	**<0.001**	**154.290**	**<0.001**
20th percentile HR	**−16.850**	**<0.001**	**−8.320**	**<0.001**	**−5.720**	**<0.001**	**135.790**	**<0.001**
80th percentile HR	**−16.720**	**<0.001**	**−7.990**	**<0.001**	**−5.750**	**<0.001**	**128.840**	**<0.001**
Mean R-R interval	**18.510**	**<0.001**	**7.970**	**<0.001**	**6.350**	**<0.001**	**148.800**	**<0.001**
Standard deviation R-R interval	**2.900**	**0.004**	1.150	0.250	1.490	0.137	**4.230**	**0.015**
Median R-R interval	**18.110**	**<0.001**	**8.120**	**<0.001**	**6.090**	**<0.001**	**144.270**	**<0.001**
Max R-R interval	**10.810**	**<0.001**	**4.580**	**<0.001**	**4.860**	**<0.001**	**56.850**	**<0.001**
Min R-R interval	**11.150**	**<0.001**	**5.050**	**<0.001**	**4.120**	**<0.001**	**56.340**	**<0.001**
20th percentile R-R interval	**17.290**	**<0.001**	**8.170**	**<0.001**	**5.940**	**<0.001**	**134.350**	**<0.001**
80th percentile R-R interval	**16.200**	**<0.001**	**7.180**	**<0.001**	**5.540**	**<0.001**	**116.690**	**<0.001**
RMSSD	**2.770**	**0.006**	1.470	0.145	1.200	0.231	**4.070**	**0.018**

Significant scores ($p < 0.05$) are shown in bold.

research.[8] Our results suggest that it is possible to detect stress using a commodity heart rate sensor, at least in the lab setting.

5.2.3 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 Welch's t-test for each feature for each pair of stressors, and one-way ANOVA using all of the stressors, in Table 12. Both tests showed 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

[8] It is important to note that we are not suggesting that the Polar H7 is better than a high-quality or clinical-grade ECG sensor. Instead, we believe with the right data processing pipeline, commodity devices might suffice for the task of stress detection.

Table 13. LOSO Cross-Validation Results for a Multiclass Classification Among Stress-Induced Periods, with the *trim_zscore* Dataset, Using Linear SVM and RF Classifiers

Prediction Metrics	SVM			RF		
	Math Class	*Book Class*	*Cold Class*	*Math Class*	*Book Class*	*Cold Class*
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

with a different label: the math test as 1, the book test as 2, and the cold test as 3. For these three-class classification tasks, we trained linear SVM and RF models for a LOSO cross validation. Table 13 shows the results. From the table, we observe that although we obtained high $F1$-scores for inferring the math and book tests, that was not the case for the cold test. In addition, although SVM and RF both produced similar prediction metrics (precision, recall, and $F1$-score), for the math and book tests, they produced widely varying results for the cold test: SVM leads to high precision with low recall, whereas RF does not show such a large difference between precision and recall. We need to look further into the modeling of different kinds of stressful periods (beyond the math, book, and cold tests discussed here) to understand this difference, which we leave to future work. Out of curiosity, we considered a two-class classification between math and book tests, and ignore the cold test completely from the evaluation (from both training and testing). We observed that the $F1$-score improved significantly for both classes, with values greater than 0.90 for both.

As in prior sections, the results here use the *trim_zscore* methods.

5.3 Evaluation in the Field Setting

In this section, we evaluate the models developed in the lab component of the study for stress detection in the field setting. As described previously, in the field component of the study, we asked the participants to wear the devices in their natural environment and prompted several EMA questions to gather the ground truth. One of those questions was "Rate your stress level over the last 10 minutes." We specifically asked about the *last 10 minutes* rather than a generic "How stressed do you feel" to reduce the errors in self-reported data due to participant recall by limiting them to think about only the last 10 minutes.

To generate the *field* dataset, we consider the physiological data collected in the 10 minutes leading to the self-report answer time. Our sampling strategy was to collect data for 1 minute every 3 minutes (i.e., sample continuously for 1 minute, then pause for 2 minutes). We took this approach to conserve battery life on the Amulet wrist devices. Hence, according to our sampling strategy, we recorded three (sometimes four) 60-second windows corresponding to the 10-minute window prior to each self-report. We computed the features for each 60-second window and labeled it as *stressed* (i.e., 1) or *not stressed* (i.e., 0) based on the response to the 5-point Likert scale report from the participant. To binarize the 5-point scale to a simple 1 or 0, for each participant we calculated the median score across all self-reports by that participant; for each report, if the score was greater than the median score, we labeled it as 1 (i.e., *stressed*), and otherwise, we labeled it as 0 (i.e., *not stressed*).

It is important to note that we evaluated the classification results (1) for the entire field data and (2) by removing the activity confounds—that is, by only considering those 60-second windows where the inferred activity level was *low*.

To infer stress in the field dataset, we used the models previously generated in the lab setting with the *trim_zscore* combination, since we achieved the highest precision, recall, and $F1$-score metrics for that condition. We used both the SVM and RF models trained on the lab dataset for classification in the field. Needless to say, the field data went through the same preprocessing methods, in this case, *trimming* outliers followed by z-score normalization. The results are shown in Table 14. We observe that removing windows with high physical activity greatly improved the prediction results, leading to a maximum $F1$-score of 0.62 when we used SVM for

Table 14. Field Evaluation for Predicting *Stress* Using SVM and RF Models Developed with the *trim_zscore* Lab Data

Prediction Metrics	*With Physical Activity*		*Removing Physical Activity*	
	SVM	RF	SVM	RF
Precision	0.46	0.42	0.58	0.49
Recall	0.57	0.55	0.67	0.61
$F1$-score	0.51	0.48	0.62	0.54

prediction. Further, by using our two-layer modeling approach, we observed that the $F1$-score improved to 0.66. Although the field $F1$-score reported by our model might seem low, it needs to be considered that we are using just a commodity device; unlike previous works that have used high-quality sensors, and fused it with other data sources like respiration (in cStress [32]) or GSR, and skin temperature (by Gjoreski et al. [25]), and attain field $F1$-scores of 0.71 and 0.63, respectively, comparable to our results.

We show that using just a commodity heart rate sensor with a rigorous data processing and feature selection pipeline, we can accurately infer stress as well as (if not better) than using an ECG device. In previous work, such as cStress, the authors [32] showed that using just the ECG data, they could infer stress in the lab with an $F1$-score of 0.78, compared to an $F1$-score of 0.87 in our case, as shown in Figure 7. They do not report field results using just the ECG sensor, so we compare our field results to a biased random classifier as the baseline. The baseline classifier randomly classifies each instance between 0 or 1, based on the probability distribution of the training set, and yields an $F1$-score of 0.44 in the field. Our approach achieves 52% better results than the baseline.

Although we show initial evidence that commodity heart rate sensors (at least the Polar H7) can be used for stress detection, there is still room for improvement. We anticipate that an increase in the sensor quality and training for a wider (and a more varied) range of stress-inducing tasks could see an increase in the inference results. In the meantime, we believe that researchers might supplement the heart rate data with other physiological data (1) to improve the accuracy or (2) to capture and compare the effect of different physiological signals in stress monitoring. Hence, it is important that our data processing pipeline works for other sensor data streams. To this end, we evaluate how well our model performs when we combine data from a GSR device to the data collected from the Polar H7.

5.4 Evaluating with GSR Data

In our study, we also asked the participants to wear a custom-made GSR sensor. We were able to record lab and field GSR data from 15 of the 27 participants in the study. We use the heart rate and GSR data from these 15 participants to build a new combined model and report the change in classification results. The GSR sensor used in this work had similar technical specifications as the one developed and evaluated by Pope et al. [44] and could measure electrodermal activity at the ventral wrist for a range of skin conductance values between 0.24 and $6.0\mu S$.

The GSR data we collected also undergoes the same rigorous data cleaning and preprocessing steps; as before, we had six different combinations of outlier handling and normalization, as shown in Table 4. Note, however, that these combinations now contain both heart rate and GSR data for 15 participants.[9] As with the heart rate data (which we now refer to as HR data), we follow a similar approach for the merged heart rate and GSR data

[9] Since the goal of this work is to evaluate how a commodity heart rate monitor works for stress measurement, we look at the combination of heart rate and GSR data. We do not report results or draw comparisons about the performance of just the GSR sensor, as it is beyond the scope of this work.

Table 15. All the Features Computed from the Filtered and Normalized GSR Data

Tonic Features	*Phasic Features*
mean of SCL (mean_SCL), max of SCL (max_SCL), min of SCL (min_SCL), and standard deviation of SCL (std_SCL)	total number of SCRs (total_SCR), sum of SCR durations (sum_dur), sum of SCR amplitudes (sum_amp), total SCR area (auc_SCR)

Table 16. Significant GSR-Based Feature Differences from the Initial Rest Period of 10 Minutes

Features	Math Test		Book Test		Cold Test	
	t-Stat	*p-Value*	*t-Stat*	*p-Value*	*t-Stat*	*p-Value*
Skin conductance mean	**−4.990**	**<0.001**	−0.450	0.651	−0.700	0.492
Skin conductance max	**−7.240**	**<0.001**	−0.860	0.390	−0.840	0.409
Skin conductance min	−1.670	0.098	−0.120	0.905	−0.330	0.742
Skin conductance standard deviation	**−5.350**	**<0.001**	−1.090	0.279	−0.780	0.442
Total number of SCRs	**−4.740**	**<0.001**	**−2.080**	**0.041**	**−2.470**	**0.022**
Sum of SCR amplitude	**−3.780**	**<0.001**	**−2.960**	**0.004**	−1.680	0.110
Sum of SCR duration	1.000	0.319	1.000	0.318	1.000	0.319
Total SCR area	0.990	0.323	1.000	0.321	1.000	0.320

Significant scores ($p < 0.05$) are shown in bold.

(which we now refer to as HR-GSR data), starting with feature computation, followed by observing significant differences, lab data results, and finally the field data results.

5.4.1 Features for HR-GSR Data. There are two main components to the overall GSR signal. The *tonic* component relates to the slower-acting components and background characteristics of the signal—that is, the overall level, slow rise, or declines over time. The common measure for the tonic component is the skin conductance level (SCL), and changes in SCL are known to reflect changes in arousal in the autonomic nervous system. For each window, we used the mean, max, min, and standard deviation of the SCL as features. The second component of the GSR signal is called the *phasic* component, which represents the faster-changing elements of the signal and is measured by skin reductance response (SCR) [11]. For each window, we compute the total number of SCRs in the window (total_SCR), sum of amplitude of the SCRs (sum_amp), sum of SCR durations (sum_dur), and the total SCR area (auc_SCR). For these latter computations, we use the EDA Explorer tool (with threshold = $0.05\mu S$) made available by Taylor et al. [52]. Table 15 lists the complete set of GSR features that, in addition to the heart rate features computed earlier, were used for the HR-GSR data.

5.4.2 Capturing Significant Difference Using GSR Data. In Section 5.1, we showed that the features computed by the heart rate sensor exhibit statistically significant differences between the baseline rest period and the stress induction periods. We also hypothesized that there might be some residual stress that is being exhibited in the initial minutes of the rest period, and by considering the last 4 minutes of the initial rest period as the baseline, we observed more features that exhibited significant difference. A similar comparison using features computed with the GSR data would help us validate whether our hypothesis was in fact true.

As in Section 5.1, we report the Welch's t-test result for the *trim_zscore* dataset. The results of the t-test using the GSR features, where we consider the entire 10 minutes of the initial rest period as the baseline, are shown in Table 16. We observe that only two and one features show significant difference for the book test and the cold test, respectively, suggesting that the GSR features are not able to capture differences between the baseline (of 10 minutes) and stress induction periods. Next we look at the t-test results by considering only the last 4 minutes of the initial rest period as the baseline (shown in Table 17). It is evident that more features exhibit significant

Table 17. Significant GSR-Based Feature Differences from the Last 4 Minutes of the Initial Rest Period

Features	Math Test		Book Test		Cold Test	
	t-Stat	*p-Value*	*t-Stat*	*p-Value*	*t-Stat*	*p-Value*
Skin conductance mean	**−3.790**	**<0.001**	−1.370	0.179	−0.860	0.393
Skin conductance max	**−5.660**	**<0.001**	**−2.010**	**0.050**	−1.340	0.187
Skin conductance min	−0.900	0.370	−0.750	0.455	0.060	0.956
Skin conductance standard deviation	**−5.970**	**<0.001**	**−2.990**	**0.004**	−2.000	0.060
Total number of SCRs	**−5.790**	**<0.001**	**−3.330**	**0.001**	**−3.920**	**<0.001**
Sum of SCR amplitude	**−4.450**	**<0.001**	**−4.580**	**<0.001**	−2.100	0.051
Sum of SCR duration	**−3.630**	**<0.001**	**−2.620**	**0.011**	**−2.400**	**0.021**
Total SCR area	**−3.670**	**<0.001**	**−3.280**	**0.001**	−2.020	0.058

Significant scores ($p < 0.05$) are shown in bold.

Table 18. LOSO Cross-Validation Results from the Different Datasets for HR-GSR Using SVM and RF, and Considering Only the Last 4 Minutes of the Rest Baseline as *Not Stressed*

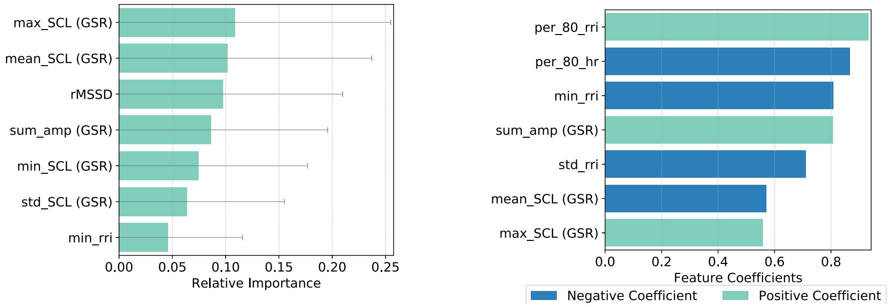
	trim_zscore		*trim_minmax*		*wins_zscore*		*wins_minmax*	
Prediction Metrics	SVM	RF	SVM	RF	SVM	RF	SVM	RF
Precision	0.86	0.92	0.83	0.86	0.85	0.87	0.79	0.77
Recall	0.89	0.91	0.86	0.84	0.82	0.88	0.82	0.79
*F*1-score	0.87	0.91	0.84	0.85	0.83	0.87	0.80	0.78

differences, which in turn supports our hypothesis about the presence of residual stress in the initial baseline period, as discussed in Section 5.1.

5.4.3 Evaluation in the Lab. To evaluate the HR-GSR datasets in the lab, we follow an approach similar to the lab evaluation in Section 5.2. In this section, we report the results for a LOSO cross validation from the different datasets, using SVM and RF, while considering only the last 4 minutes of the initial rest period as *not stressed*. The results obtained are shown in Table 18. We observe an increase in the *F*1-score, once we include the GSR data. Although the best results obtained were from *trim_zscore* (as for HR-only data), it is interesting to see that an RF model performed better than SVM for HR-GSR data (whereas SVM was better for HR data). Next, we used this RF model with our proposed two-layer model and observed that the *F*1-score improved to 0.94.

As for HR-only data, we report the feature importance results for the HR-GSR data in Figure 8. From the feature importance plots, we observe that the heart rate features (obtained from a commodity sensor) did play an important role in the overall classification model, even when combined with a custom sensor, which suggests that using a custom sensor does not obviate the need for the heart rate sensor.

5.4.4 Evaluation in the Field. As in the field evaluation of the HR-only data, we use the model built during the lab evaluation of the HR-GSR data to predict the stress labels in the field study. We report the results in Table 19. We observe that combination of GSR data and HR data results in an *F*1-score of 0.70, which improves to 0.73 when considered in conjunction with the Bayesian network model in our proposed two-layer approach. Based on this result, it *seems* that *trim_zscore* can be used as a standard data processing step for accurately detecting stress from a commodity heart rate monitor, and it can be extended to other sensor streams being used in combination (GSR in our case), with similar levels of classification performance as the clinical-grade or custom sensor–based systems used by other researchers in the past.



(a) Feature importance using RF: The green bars represent the feature importance in the forest, along with their intertree variability.

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

Fig. 8. Feature importance representation using RF and linear SVM, using the *trim_zscore* combination of HR-GSR data, sorted from highest to lowest. For the sake of space, we only show the top seven features.

Table 19. Field Evaluation for Predicting *Stress* Using SVM and RF Models Developed with the *trim_zscore* HR-GSR Lab Data

Prediction Metrics	*With Physical Activity*		*Removing Physical Activity*	
	SVM	RF	SVM	RF
Precision	0.53	0.52	0.57	0.57
Recall	0.81	0.87	0.83	0.91
$F1$-score	0.64	0.65	0.68	0.70

6 DISCUSSION AND FUTURE WORK

Although 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.

Capturing context. Currently, we infer stress with data from a commercial heart rate sensor. However, researchers have shown that contextual information can also be useful for inferring stress [37]. With the popularity of smartphones and smartwatches, and with the availability of multiple sensors on these devices, context monitoring has become relatively straightforward. In the future, we intend to augment the data obtained from the heart ate sensor with contextual information to identify stress and nonstress periods. If we find that contextual markers help in identifying stress, context may also be used to predict a stressful situation before it occurs. This, in turn, will open new research directions. In our work, we use one aspect about the context of the user—*activity*, which helps us remove instances in the physiological signals that could be caused by physical activity. Although we use an Amulet to measure activity and administer EMAs, we anticipate that other mobile and wearable devices may also be suitable for detecting motion or physical activity context.

Capturing other types of stressors. In this work, through extensive lab studies, we looked at three specific types of stressors: the mental arithmetic test, startle response test, and ice water test. Although these three stressors capture a variety of physiological responses from the participants, we believe that conducting laboratory tests and adapting our models with a wider range of real-life situations (e.g., social interactions, written arithmetic, eating, public speaking) will lead to a more detailed mapping of the "stresses" a person goes through in a day.

The additional physiological responses, coupled with the context information mentioned earlier, should help advance the research in making fine-grained inferences about a person's stress levels.

Standardizing the processing pipeline. Prior research using clinical-grade sensors to monitor stress indicates that winsorization is an effective preprocessing approach to handling outliers [26, 32]. For commodity, off-the-shelf sensors (Polar H7), we found that the results obtained after performing simple *trimming* outperformed the results obtained after performing winsorization. This observation indicates that data preprocessing and data cleaning techniques that perform well for clinical-grade devices might not perform equally well for commodity devices. One possible explanation for the difference might be the noise present in the data. Since clinical-grade sensors are more robust than their consumer-grade counterparts, we expect that they are less prone to environmental noise. It may be that the outliers present in the data obtained from such clinical devices might have some useful information, making winsorization an effective method. This example is one that could be directly compared to previous works; there might be more such instances where choices for data processing and data modeling might lead to differing performance in commodity and clinical devices. Based on our analyses, however, it seems that *trimming* along with *z-score* normalization might be a better way to process data for the task of stress detection. The proposed steps led to the best classification results for both HR-only data and HR-GSR data, coming from commodity and custom-made devices. This suggests that our processing pipeline is robust for two device types and may be appropriate for other sensor streams as well.

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 the tasks was the same for all participants: the math test, followed by the book drop test, and finally the cold water test. Although the decision to follow the same sequence of tasks is consistent with previous works [32, 42], some researchers claim that randomizing the order may be required to avoid a carry-over effect of previous tasks. Further exploration is required to observe whether and how the order affects the results.

Usability of chest-based sensors. Some believe that usually chest-band-based sensors might be bulky and uncomfortable, making them a poor choice for continuous usage. We chose to use the Polar H7, a very popular product, with more than 2,300 five-star reviews on Amazon.[10] Further, in our study, after the field session was concluded, we conducted a quantitative and qualitative survey about how the participants felt about our data collection system. When asked if they "found the system to cause physical discomfort," participants mostly disagreed (mean = 1.22 on a scale from 0 to 4, with 0 being strongly disagreed and 4 being strongly agreed); only three participants agreed that the device caused discomfort. On being asked if they "could have worn the system for a longer period of time," participants mostly agreed (mean = 2.81). In fact, participants were more concerned about the comfort factor of the wrist-based GSR device. Almost 50% of the participants noted they found the GSR device to cause physical discomfort—we believe that this was because of the protruding electrodes.

It is important to note that we surveyed college students who wore the sensors for 3 days and only during waking hours. And although the participants agreed that they could have worn the sensors for longer, the comfort and acceptability of actual long-term usage might vary. It is possible, given the demographics and duration of the study, that the usability results might be optimistically biased. Further long-term usability tests and surveys, across demographics, need to be conducted.

Scalability of the stress detection model. To observe the sensitivity of our model on the number of users for training the model, we modeled performance for different subsets of users. Considering n to be the number of users in the model, where $n \in [3, 26]$, all possible combinations of users for each n is ${}^{26}C_n$. For each n, we randomly selected 200 combinations[11] 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[12] in Figure 9.

[10] https://www.amazon.com/Polar-Bluetooth-Sensor-Fitness-Tracker/dp/B00NOHWTO6/ref=cm_cr_arp_d_product_top?ie=UTF8.

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

[12] The results reported are using the *trim_zscore* combination of the lab HR data.

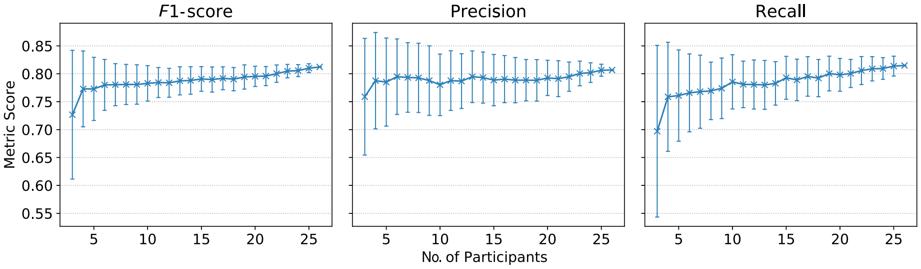


Fig. 9. Performance of LOSO cross validation for detecting stress periods using n number of participants, where $n \in [3, 26]$. The points represents the mean of the 200 randomly selected combinations of participants, and the error bars represent the standard deviation across the different combinations.

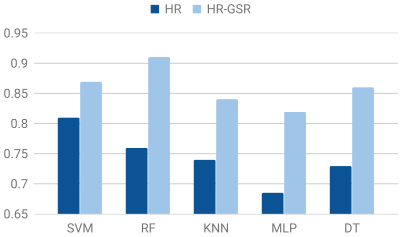


Fig. 10. Comparison of different algorithms for detecting stress in the lab using *trim_zscore* data for both HR only and HR-GSR.

It is interesting to observe that the metrics, the $F1$-score for example, vary little after 15 or 16 participants, with substantial reduction in the standard deviation. Although we realize that the convergence at $n = 26$ is because there is only one combination, it is reassuring to see that the model performance does not vary substantially with different participant combinations. This result is, of course, just for a college student population; we plan to evaluate this issue in greater detail for a broader population in lab and field situations in future work.

It is important to note that our current work (along with prior works in this domain) is on *offline* stress detection. Our methods leverage the entire dataset for each participant and retrospectively infer stressful events. This approach helps us effectively normalize the participant's physiological signals because we can easily compute metrics such as mean, standard deviation, min, and max over the data. Normalization over continuously streaming data would be a nontrivial task. One approach would be to calculate local normalization over a sliding window; another approach is for the system to observe a person's physiological signals for a brief initial period before calculating parameters for normalization and subsequent stress detection.

Comparison with other algorithms. In our work, we focused on two base classifiers: SVM and RF. Gjoreski et al. [26] compared performance of several classifiers, including SVM, RF, KNN, and decision tree, and found that SVM and RF outperformed all other algorithms. We also made a similar comparison with three additional algorithms: KNN, multilayer perceptron with ReLU, and decision tree. We compare the performance of the different algorithms for both the HR-only data and HR-GSR data in the lab and show the comparison in Figure 10. Since we were interested in the best-performing classifier, we report results only from the classifier and not when combined with the Bayesian network model in our proposed two-layer approach. We observe from Figure 10 that SVM and RF are the top-performing classifiers for both HR-only and HR-GSR data, similar to findings by Gjoreski et al. [26].

Need for further multiscale deployment and evaluation. In our work, we looked at the performance of the Polar H7 on 26 participants, in the lab and over 3 days of *free-living* situations. We believe that further research with more participants and for a longer duration 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 that the use of cheap commodity devices will help overcome these shortcomings of "unavailable" devices.

7 CONCLUSION

Unlike current approaches for measuring an individual's mental stress through the use of expensive clinical-grade sensors, this article explores the possibility of using a cheap commercial wearable device to identify whether an individual is undergoing mental stress in the lab and in free-living settings. As the first work testing the viability of commercial sensors for identifying stress, especially in a field setting, we demonstrated the potential of this approach. We also identified several important methodological issues worthy of future work. After thorough data preprocessing and cleaning steps, we found that techniques reported to perform well for clinical-grade devices seem to underperform for consumer-grade devices, as in the example of winsorization. We further proposed a novel two-layer approach for detecting stress by accounting for the stress in the previous time window.

For the lab study, using just a commodity heart rate monitor, we found that we could identify stressful periods with an $F1$-score of 0.87. Moreover, in the best case, we could distinguish between three different types of stress-inducing tasks with an $F1$-score of 0.82. In the future, researchers should look beyond distinguishing "stressful" periods from "nonstressful" periods to identifying the type of stress a person is currently undergoing, to enable adaptive just-in-time interventions. Additionally, in the lab setting, we found that augmenting the heart rate sensor with a GSR sensor helped boost the $F1$-score from 0.87 to 0.94. For the field study, we found that we could identify stress with an $F1$-score of 0.66 using heart rate data alone and one of 0.72 using heart rate and GSR data together. In the future, we plan to use contextual information to improve the stress detection performance.

REFERENCES

[1] M. Al'Absi and D. K. Arnett. 2000. Adrenocortical responses to psychological stress and risk for hypertension. *Biomedicine & Pharmacotherapy* 54, 5 (2000), 234–244.

[2] Mustafa Al'Absi, Stephan Bongard, Tony Buchanan, Gwendolyn A. Pincomb, Julio Licinio, and William R. Lovallo. 1997. Cardiovascular and neuroendocrine adjustment to public speaking and mental arithmetic stressors. *Psychophysiology* 34, 3 (May 1997), 266–275. DOI: https://doi.org/10.1111/j.1469-8986.1997.tb02397.x

[3] Marco Altini. 2015. Hardware for HRV: What Sensor Should You Use? Retrieved February 18, 2020 from https://www.hrv4training.com/blog/hardware-for-hrv-what-sensor-should-you-use.

[4] Android Central. 2016. Heart Rate Monitor Horribly Inaccurate? Retrieved February 18, 2020 from https://forums.androidcentral.com/samsung-gear-2-gear-2-neo/380007-heart-rate-monitor-horribly-inaccurate.html.

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

[6] George E. Billman. 2013. The LF/HF ratio does not accurately measure cardiac sympatho-vagal balance. *Frontiers in Physiology* 4 (2013), 26. DOI: https://doi.org/10.3389/fphys.2013.00026

[7] Biopac Systems Inc.2016. Biopac MP150. Retrieved February 18, 2020 from https://www.biopac.com/wp-content/uploads/MP150-Systems.pdf.

[8] George Boateng, Ryan Halter, John A. Batsis, and David Kotz. 2017. ActivityAware: An app for real-time daily activity level monitoring on the amulet wrist-worn device. In *Proceedings of the IEEE PerCom Workshop on Pervasive Health Technologies (PerHealth'17)*. IEEE, Los Alamitos, CA, 431–435. DOI: https://doi.org/10.1109/PERCOMW.2017.7917601

[9] George Boateng, Vivian Genaro Motti, Varun Mishra, John A. Batsis, Josiah Hester, and David Kotz. 2019. Experience: Design, development and evaluation of a wearable device for mHealth applications. In *Proceedings of the International Conference on Mobile Computing and Networking (MobiCom'19)*. DOI: https://doi.org/10.1145/3300061.3345432

[10] Andrey Bogomolov, Bruno Lepri, Michela Ferron, Fabio Pianesi, and Alex Sandy Pentland. 2014. Pervasive stress recognition for sustainable living. In *Proceedings of the IEEE International Conference on Pervasive Computing and Communication Workshops, (PerCom Workshops'14)*. IEEE, Los Alamitos, CA, 345–350. DOI: https://doi.org/10.1109/PerComW.2014.6815230

[11] J. J. Braithwaite, D. G. Watson, R. Jones, and M. Rowe. 2013. A guide for analysing electrodermal activity (EDA) & skin conductance responses (SCRs) for psychological experiments. *Psychophysiology* 49, 8 (2013), 1–42.

[12] Grace Chen, Varun Mishra, and Ching-Hua Chen. 2019. Temporal factors of listening to music on stress reduction. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers (UbiComp/ISWC'19 Adjunct)*. ACM, New York, NY, 907–914. DOI: https://doi.org/10.1145/3341162.3346272

[13] Jongyoon Choi, B. Ahmed, and Ricardo Gutierrez-Osuna. 2012. Development and evaluation of an ambulatory stress monitor based on wearable sensors. *IEEE Transactions on Information Technology in Biomedicine* 16, 2 (2012), 279–286. DOI: https://doi.org/10.1109/TITB.2011.2169804

[14] George P. Chrousos and Philip W. Gold. 1992. The concepts of stress and stress system disorders: Overview of physical and behavioral homeostasis. *JAMA* 267, 9 (1992), 1244–1252.

[15] Matteo Ciman, Katarzyna Wac, and Ombretta Gaggi. 2015. iSenseStress: Assessing stress through human-smartphone interaction analysis. In *Proceedings of the International Conference on Pervasive Computing Technologies for Healthcare (PerHealth'15)*. 84–91.

[16] Sheldon Cohen, Tom Kamarck, and Robin Mermelstein. 1983. A global measure of perceived stress. *Journal of Health and Social Behavior* 24, 4 (1983), 385–396.

[17] Brian Edwards. 2011. Wearable Sensor by Affectiva Can Measure Anxiety and Is Helping Autism Research. Retrieved February 18, 2020 from https://www.imedicalapps.com/2011/10/wearable-sensor-by-affectiva-can-measure-anxiety-and-is-helping-autism-research/.

[18] 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 *Proceedings of the IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops'17)*. IEEE, Los Alamitos, CA, 673–678. DOI: https://doi.org/10.1109/PERCOMW.2017.7917644

[19] Emaptica. 2018. Empatica E4. Retrieved February 18, 2020 from https://www.empatica.com/en-eu/research/e4/.

[20] 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 (Sept. 2014), 535–541. http://www.ncbi.nlm.nih.gov/pubmed/25177179.

[21] S. M. Fox and W. L. Haskell. 1970. The exercise stress test: Needs for standardization. In *Cardiology: Current Topics and Progress*. Academic Press, New York, NY, 149–154.

[22] Ulla Fredriksson-Larsson, Eva Brink, Gunne Grankvist, Ingibjörg H. Jonsdottir, and Pia Alsen. 2015. The single-item measure of stress symptoms after myocardial infarction and its association with fatigue. *Open Journal of Nursing* 5, 04 (2015), 345.

[23] 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*. DOI: https://doi.org/10.4108/icst.mobihealth.2014.257418

[24] Enrique Garcia-Ceja, Venet Osmani, and Oscar Mayora. 2016. Automatic stress detection in working environments from smartphones' accelerometer data: A first step. *IEEE Journal of Biomedical and Health Informatics* 20, 4 (July 2016), 1053–1060. DOI: https://doi.org/10.1109/JBHI.2015.2446195

[25] 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: Adjunct (UbiComp'16 Adjunct)*. ACM, New York, NY, 1185–1193. DOI: https://doi.org/10.1145/2968219.2968306

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

[27] Kristina Grifantini. 2010. Sensor Detects Emotions Through the Skin. Retrieved February 18, 2020 from https://www.technologyreview.com/s/421316/sensor-detects-emotions-through-the-skin/.

[28] Tian Hao, Kimberly N. Walter, Marion J. Ball, Hung-Yang Chang, Si Sun, and Xinxin Zhu. 2017. StressHacker: Towards practical stress monitoring in the wild with smartwatches. In *Proceedings of the AMIA Annual Symposium*.

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

[30] Javier Hernandez, Rob R. Morris, and Rosalind W. Picard. 2011. *Call center stress recognition with person-specific models*. In *Affective Computing and Intelligent Interaction*. Lecture Notes in Computer Science, Vol. 6974. Springer, 125–134 pages. DOI: https://doi.org/10.1007/978-3-642-24600-5_16

[31] Josiah Hester, Travis Peters, Tianlong Yun, Ronald Peterson, Joseph Skinner, Bhargav Golla, Kevin Storer, et al. 2016. Amulet: An energy-efficient, multi-application wearable platform. In *Proceedings of the ACM Conference on Embedded Networked Sensor Systems (SenSys'16)*. ACM, New York, NY, 216–229. DOI: https://doi.org/10.1145/2994551.2994554

[32] 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. *In Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp'15)*. 493–504. DOI: https://doi.org/10.1145/2750858.2807526

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

[34] 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. *Frontiers in Psychology* 8 (2017), 213. DOI: https://doi.org/10.3389/fpsyg.2017.00213

[35] Hong Lu, Denise Frauendorfer, Mashfiqui Rabbi, Marianne S. Mast, Gokul T. Chittaranjan, Andrew T. Campbell, Daniel G. Perez, and Tanzeem Choudhury. 2012. StressSense: Detecting stress in unconstrained acoustic environments using smartphones. In *Proceedings of the ACM Conference on Ubiquitous Computing (UbiComp'12)*. DOI: https://doi.org/10.1145/2370216.2370270

[36] Bruce S. McEwen and Eliot Stellar. 1993. Stress and the individual: Mechanisms leading to disease. *Archives of Internal Medicine* 153, 18 (1993), 2093–2101.

[37] Varun Mishra, Tian Hao, Si Sun, Kimberly N. Walter, Marion J. Ball, Ching-Hua Chen, and Xinxin Zhu. 2018. Investigating the role of context in perceived stress detection in the wild. In *Proceedings of the 2018 ACM International Joint Conference and the 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers (UbiComp'18)*. ACM, New York, NY, 1708–1716. DOI: https://doi.org/10.1145/3267305.3267537

[38] 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 *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp'18)*. ACM, New York, NY, 1717–1728. DOI: https://doi.org/10.1145/3267305.3267538

[39] Amir Muaremi, Bert Arnrich, and Gerhard Tröster. 2013. Towards measuring stress with smartphones and wearable devices during workday and sleep. *BioNanoScience* 3, 2 (2013), 172–183.

[40] 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 *Proceedings of the IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI'14)*. IEEE, Los Alamitos, CA, 185–188. DOI: https://doi.org/10.1109/BHI.2014.6864335

[41] NIMH. 2016. 5 Things You Should Know About Stress. Retrieved February 18, 2020 from https://www.nimh.nih.gov/health/publications/stress/index.shtml.

[42] K. Plarre, A. Raij, S. M. Hossain, A. A. Ali, M. Nakajima, M. Al'Absi, E. Ertin, et al. 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'11)*. IEEE, Los Alamitos, CA, 97–108. http://ieeexplore.ieee.org/xpls/abs_all.jsp? arnumber=5779068.

[43] Polar. 2017. Polar H7. Retrieved February 18, 2020 from https://support.polar.com/us-en/support/H7_heart_rate_sensor

[44] Gunnar C. Pope, Varun Mishra, Stephanie Lewia, Byron Lowens, David Kotz, Sarah Lord, and Ryan Halter. 2018. An ultra-low resource wearable EDA sensor using wavelet compression. In *Proceedings of the IEEE Conference on Body Sensor Networks (BSN'18)*. 193–196. DOI: https://doi.org/10.1109/BSN.2018.8329691

[45] R. Rosmond and P. Björntorp. 1998. Endocrine and metabolic aberrations in men with abdominal obesity in relation to anxio-depressive infirmity. *Metabolism* 47, 10 (1998), 1187–1193.

[46] Pedro Sanches, Kristina Höök, Elsa Vaara, Claus Weymann, Markus Bylund, Pedro Ferreira, Nathalie Peira, and Marie Sjölinder. 2010. Mind the body! Designing a mobile stress management application encouraging personal reflection. In *Proceedings of the ACM Conference on Designing Interactive Systems*. ACM, New York, NY, 47–56.

[47] Akane Sano and Rosalind W. Picard. 2013. Stress recognition using wearable sensors and mobile phones. In *Proceedings of the Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, Los Alamitos, CA, 671–676. DOI: https://doi.org/10.1109/ACII.2013.117

[48] Hillol Sarker, Inbal Nahum-Shani, Mustafa Al'Absi, Santosh Kumar, Matthew Tyburski, Md M. Rahman, Karen Hovsepian, et al. 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'16)*. ACM, New York, NY, 4489–4501. DOI: https://doi.org/10.1145/2858036.2858218

[49] Fred Shaffer and J. P. Ginsberg. 2017. An overview of heart rate variability metrics and norms. *Frontiers in Public Health* 5 (2017), 258. DOI: https://doi.org/10.3389/fpubh.2017.00258

[50] Feng-Tso Sun, Cynthia Kuo, Heng-Tze Cheng, Senaka Buthpitiya, Patricia Collins, and Martin Griss. 2012. Activity-aware mental stress detection using physiological sensors. In *Mobile Computing, Applications, and Services*. Lecture Notes of the Institute for Computer Sciences, Social Informatics, and Telecommunications Engineering, Vol. 76. Springer, 211–230. DOI: https://doi.org/10.1007/978-3-642-29336-8_16

[51] 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), 153–156. http://www.ncbi.nlm.nih.gov/pubmed/11153730.

[52] Sara Taylor, Natasha Jaques, Weixuan Chen, Szymon Fedor, Akane Sano, and Rosalind Picard. 2015. Automatic identification of artifacts in electrodermal activity data. In *Proceedings of the 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'15)*, Vol. 2015. 1934–1937. DOI: https://doi.org/10.1109/EMBC.2015.7318762

[53] Julian F. Thayer, Fredrik Åhs, Mats Fredrikson, John J. Sollers, and Tor D. Wager. 2012. A meta-analysis of heart rate variability and neuroimaging studies: Implications for heart rate variability as a marker of stress and health. *Neuroscience and Biobehavioral Reviews* 36, 2 (Feb. 2012, 747–756. DOI: https://doi.org/10.1016/j.neubiorev.2011.11.009

[54] Vincent W. S. Tseng, Saeed Abdullah, Min Hane Aung, Franziska Wittleder, Michael Merrill, and Tanzeem Choudhury. 2016. Assessing mental health issues on college campuses: Preliminary findings from a pilot study. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct (UbiComp'16 Adjunct)*. ACM, New York, NY, 1200–1208. DOI:https://doi.org/10.1145/2968219.2968308

[55] Rui Wang, Peilin Hao, Xia Zhou, Andrew T. Campbell, and Gabriella Harari. 2015. SmartGPA: How smartphones can assess and predict academic performance of college students. *GetMobile: Mobile Computing and Communications* 19, 4 (Oct. 2015), 13–17. DOI:https://doi.org/10.1145/2904337.2904343

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

[57] Zephyr. 2018. Zephyr BioHarness 3. Retrieved February 18, 2020 from https://www.zephyranywhere.com/media/download/bioharness3-user-manual.pdf.

[58] Amazon. 2020. BioHarness 3 - Wireless Professional Heart Rate & Physiological Monitor with Bluetooth. https://www.amazon.com/BioHarness-Wireless-Professional-Physiological-Bluetooth/dp/B009ZUYNCW.

Received November 2018; revised July 2019; accepted September 2019