Published in final edited form as:

Proc ACM Int Conf Ubiquitous Comput. 2015 September ; 2015: 493–504. doi: 10.1145/2750858.2807526.

cStress: Towards a Gold Standard for Continuous Stress Assessment in the Mobile Environment

Karen Hovsepian[ᵛ], **Mustafa al'Absi**[†], **Emre Ertin**[ˢ], **Thomas Kamarck**[^], **Motohiro Nakajima**[†], and **Santosh Kumar**[★]

[ᵛ]Troy University

[†]University of Minnesota Medical School

[ˢ]The Ohio State University

[^]University of Pittsburgh

[★]University of Memphis

Abstract

Recent advances in mobile health have produced several new models for inferring stress from wearable sensors. But, the lack of a gold standard is a major hurdle in making clinical use of continuous stress measurements derived from wearable sensors. In this paper, we present a stress model (called *cStress*) that has been carefully developed with attention to every step of computational modeling including data collection, screening, cleaning, filtering, feature computation, normalization, and model training. More importantly, cStress was trained using data collected from a rigorous lab study with 21 participants and validated on two independently collected data sets — in a lab study on 26 participants and in a week-long field study with 20 participants. In testing, the model obtains a recall of 89% and a false positive rate of 5% on lab data. On field data, the model is able to predict each instantaneous self-report with an accuracy of 72%.

Author Keywords

Stress; mobile health (mHealth); wearable sensors; modeling

INTRODUCTION

Thanks to advances in diagnostic and analytical methods of modern medicine, we are beginning to more clearly see the large role that excessive and/or lingering psychological stress plays in the decline of our emotional and physical well-being, being implicated in such illnesses as diabetes, depression, heart diseases, and digestive problems [43, 26, 27, 11,

3, 4, 18, 13, 14]. In addition to long term negative effects on health, stress may also cause flare-ups in those suffering from migraines or other stress disorders [44]. A timely intervention to manage daily stress can significantly improve our physical, physiological, psychological, behavioral, and social health.

A sensor-based continuous measurement of stress in daily life has a potential to increase awareness of patterns of stress occurrence, its antecedents, and its precipitants [47]. Fortunately, wearable sensors have progressed to the point that they can continuously measure physiology and wirelessly stream the data to a smartphone for real-time analysis. This, coupled with computational modeling advances, has led to several recent works on continuous measurement of stress in the mobile environment [22, 36, 2].

Despite these advances, we still lack a well-validated stress model that can be used for clinically managing daily stress in the natural environment. There are several challenges in developing and validating such a model. First, there is no universally-accepted definition of stress. Second, there is no gold standard, either in the lab or in the field. For example, cortisol (measured from blood or saliva samples) is often referred to as stress hormone and self-reporting is the most commonly used method to assess stress in the field. But, correlations between cortisol and self-reports have been limited to 0.26–0.36 [5, 6]. Third, physiological data collected from the field are subject to numerous sources of noise and losses [38]. For example, when sensors such as electrocardiogram (ECG) are worn throughout the day, their attachment with the skin can degrade. Physical movements could also introduce noise in the data due to jerks to electrodes. Data could get lost in wireless transmission.

The fourth major challenge is dealing with confounding variables. Physiological arousal that should be indicative of a stress response can be easily obfuscated by movements of limbs, changes in posture, and physical activity. Separating out good quality data that can be analyzed to determine whether it represents a stress-response is therefore a significant challenge.

The fifth challenge is identification and computation of discriminative features that can identify and distinguish a stress response from other similar physiological arousals. Finally, developing a computational model from these features and training and validating it for field usage is another significant challenge, especially due to lack of a gold standard with which to train or validate the model.

For a model to be considered a gold standard for continuous stress assessment, we consider the following two criteria —1.) reproducibility (i.e., validity) of the model on an independently collected dataset in both lab and field studies; 2.) high accuracy, i.e., a recall rate ≥90% with a false positive rate of under 5% on independently collected lab data (when using lab protocol as the label) and an accuracy of ≥70% in the field setting (when using self-report as the label). We note that when comparing self-report items for consistency (using Cronbachs Alpha measure), 0.7 is the rule-of-thumb threshold for declaring a concordance [33]. This threshold reflects inherent variabilities and biases in self-report data.

In this paper, we address each of the above six challenges and present the *cStress* model, which represents a solid step towards establishing a gold standard for continuous stress assessment. *cStress* analyzes a minute's worth of ECG and respiration data[1], and, if this minute is not confounded by physical activity, outputs probability of stress. It is trained using data collected from 21 participants who were subjected to three validated stressors — public speaking, mental arithmetic, and cold-pressor. The ground-truth in the lab is collected for each minute, based on knowledge of starts and ends of simulated stressors. This enabled us to create a fine-grained model of physiological stress activation (at one minute resolution). The model is evaluated also at the same fine-grained level, i.e. once a minute on the lab data. In the field, self-reported stress in response to Ecological Momentary Assessment (EMA) prompts are used as ground-truths. We note that even though *cStress* produces a stress value for each minute, participants were prompted for self-report of stress only 15 times a day [45] (in order to limit participant fatigue). Field validation of *cStress* is, therefore, limited to these self-reports.

The *cStress* model achieves a recall (true positive) rate of 88.6% and a false positive rate of 4.65% on (1,501 minutes of) test dataset from the lab. When the output of *cStress* is compared with each of the 14 self-reports from each participant (in the lab session) to obtain an accuracy value for each participant, we report a median accuracy of 90%. When comparing with each (of the 1,060) self-report collected in the field (consisting of 1,000+ hours of sensor data), we obtain a median accuracy of 72%. We also rank the features and find that 80th percentile and mean of interbeat interval (i.e., time between successive R peaks in ECG) and mean and median of ratio between inspiration and expiration duration in respiration are the most informative features.

MODELING OVERVIEW

Figure 1 presents an overview of the *cStress* model, starting from data processing and culminating with the training and validation process. The entire model is built using data collected via a robust wearable sensor suite, called AutoSense [15], which we describe in further detail in the next section. AutoSense sensors are used to collect the physiological data in both lab and field.

The lab study data are collected at the University of Minnesota Medical School, using a carefully designed lab study protocol. They are used to train and validate the model using labels (i.e., ground truth or gold standard) constructed from the lab protocol coding. The minutes of the lab session during which a participant undergoes a stress protocol are considered to be in the 'stressed' class, and 'not stressed' otherwise. This is similar to the approach followed in [36].

Field data are collected at the University of Memphis, and are used to validate the model in the participants' natural environment. In this case, validation ground-truth is based on self-

[1]We also note that the use of ECG and respiration sensors by our model is well founded on extensive prior research on physiological responses to changes in stress and emotion [15]. While there are other physiological manifestations of stress, such as changes in skin conductivity, skin temperature, and blood-pressure, ECG and respiration are the primary ones [17, 20, 21, 40].

reports filled-out at random times throughout the day, which assess the participants' stress state at the time of each prompt.

The first step in constructing the *cStress* model is to assign correct time-stamps to the data received over the wireless channel from wearable sensors. For time synchronization across all measurements collected from wearable sensors and the phone, data is time-stamped when it is received at the phone. Data losses and software delays on the phone introduce variability in the time-stamping process. The granularity of *cStress* is at the level of a minute while the errors in timestamps may be on the order of milliseconds since the data is transmitted tens of times each second. The main issue of time synchronization occurs due to data loss. Time-stamp calibration is, therefore, needed to distinguish packet delays from packet losses. Once we determine that packets are lost, we can take corrective actions (e.g., interpolations). To do time-stamp calibration, we developed a dynamic programming algorithm to infer the correct time-stamp of each received data sample and identify the lost data samples.

Second, we interpolate any lost data if the loss is minimal so as not to degrade the overall data quality. The third step is to identify and screen out poor quality data that can lead to erroneous inferences. Rigorous data processing is essential to obtain usable results from physiological data collected in the field, due to the expected presence of noise and artifacts. The major causes of data degradations and losses in sensor measurements (e.g., attachment loosening, physical movements, etc.) are analyzed in detail in [38], which found that data yield using AutoSense is better compared to other previously reported field studies using wireless physiological sensors. The fourth step is to detect physical activity and exclude corresponding data from the application of the *cStress* model.

Data remaining after the above steps are used to compute a variety of base features from both ECG and respiration. The features are subsequently screened to remove any remaining outliers (e.g., long beat-to-beat interval in ECG due to a missed or spurious beat). To reduce participant dependency and make the model generalizable, the training features should not exhibit any participant-specific effects, such as participant-specific mean and standard-deviation. Therefore, a critical step in pre-processing is the normalization of each lab-study participant's features. Further, normalization is also carried out for any subsequent participant on whom the model is applied. We introduce two ideas for robust normalization. The first is to use a technique called winsorization [50] to limit the impact of any outliers and the second is to compute the overall mean and standard deviation only from those data that are not affected by intense physical activity, which significantly deviate from baseline.

The normalized features are aggregated into one-minute blocks/windows, by computing various statistical features (e.g. average, variance) per block. The one-minute granularity has been the standard in lab and ambulatory physiological monitoring [16, 17, 20, 21, 22, 36, 2] because this level of aggregation allows relatively robust and stable feature statistics. Using blocks of less than 1 minute increases variability, which may lead to degraded model performance.

We next use the aggregated normalized features, representing each one-minute block, to train a Support Vector Machine (SVM) [8] model and optimize its hyper-parameters to

maximize the F1 score. The SVM algorithm has been shown to have a comparable or better performance (compared to other machine learning models) for inferring stress on a minute-by-minute basis [36]. For training the model, we use cross-subject validation on the training data to optimize the training algorithm's hyper-parameters. During all subsequent validations/applications of the model, we apply the model on each participant separately. For validation on field data, we develop a Bayesian Network (BN) model that uses *cSress* to infer the instantaneous self-reports, used as field ground-truth. The use of a BN helps to address the arbitrary lags between physiological response to a stressor and its memory in the mind, which is captured in self-reports.

Code Release

The source code for *cSress* will be released as open source software via the MD2K Center of Excellence[2].

DATA COLLECTION

To train and test the *cStress* model, we use sensor and self-report data collected in three user studies — two lab studies (with $n = 24$ and $n = 26$) and a field study with $n = 30$. Data from the first lab study, which we refer to as ***train***, is used to train and cross-subject-validate *cStress*. The second lab study is referred to as ***test*** and is used for out-of-sample testing of *cStress*. The third dataset is called ***field***. This data is used to validate *cStress* in the much noisier real-life conditions against self-reported stress. We now describe the devices, participants, study procedure, and the collected data.

Devices and Sensing Modalities

During the study period, participants wore a sensor suite underneath their clothes with similar functionality as BioHarness [1]. The sensor suite, called AutoSense [15], consists of several biomedical sensors. These include an unobtrusive, flexible band worn around the chest, providing respiration data by measuring the expansion and contraction of the chest via inductive plethysmography (RIP), a two-lead electrocardiograph (ECG) and 3-axis accelerometers.

The measurements collected by the sensors are transmitted wirelessly, using ANT radio, to an Android smart phone. The sampling rates for the sensors are 128 Hz for ECG, 21.3 Hz for respiration, and 16 Hz for each accelerometer axis. These samples were transmitted at the rate of 28 packets/second, where each packet contains 5 samples. Each participant also carried a smart phone that received and stored data transmitted by the sensor suite and collected self-reports. The sensors last around 10 days between successive battery recharges.

Lab-study Data

We follow the same protocol for the lab study as reported in [36]. Participants were asked to sit in a comfortable chair and rest for 30 minutes during the initial baseline. Three types of validated stressors — socioevaluative, cognitive, and physical challenges were used. During

[2]See the website of the NIH Center of Excellence for Mobile Sensor Data to Knowledge (MD2K): https://md2k.org.

the socioevaluative challenge, the participant was given a topic and asked to prepare (for 4 minutes) and deliver (for 8 minutes) a speech in front of a research staff. For a cognitive challenge (4 minutes), the participant was given a three digit number and asked to add three digits of that number, and then add the sum to the three digit number. Participants in the ***train*** study repeated this while seated and standing (counterbalanced). Participants in the ***test*** session completed only a single instance of this task while being seated (because no significant effect of change in posture on stress response was observed in the ***train*** dataset). Finally, during the physical stressor, the participant was asked to leave his/her hand submerged in ice cold water, for 90 seconds. This was followed by a 30-minute rest period to allow the participants' physiology and mental state to return to baseline.

These tasks have been shown to reliably induce stress-related physiological changes [5]. Therefore, the lab protocol is used as a gold standard during the lab study rather than using self-reports. Time-stamping each distinct rest and stress period allows us to construct ground-truth labels for each minute of the lab-session, designating a minute as stressed (class 1) if the participant was undergoing a stress task during that minute, or not-stressed (class −1) otherwise. These labels are subsequently used to train the *cStress* model.

Field data

For the field study, 23 participants wore the sensors for seven days in their natural field environment. They were instructed to wear the sensors during their entire waking hours (lasting approximately 10–16 hours each day). They reported to the lab each day to verify the functioning of the sensors. The data quality was also assessed continuously by the smartphone; the status of both data quality and wireless connection status with the sensors was displayed (similar to the status of Wi-Fi signal strength icon). Participants were prompted to fix the attachment or wireless connection if good quality data was not received. They were instructed on how to fix both of these at the time of their recruitment.

Self-reports

Participants in the lab session were asked to provide self-reported stress level in the lab 14 times, including before and after each stress session. During the field study, participants are prompted an average of 15 times daily, at random times (and sometimes in response to self-reports of smoking and alcohol use) to answer a questionnaire that constitutes an Ecological Momentary Assessment (EMA).

A self-report of stress in both lab and field contains five questions — "Cheerful?", "Happy?", "Angry/Frustrated?", "Nervous/Stressed?", and, "Sad?". These five items represent an adaptation of the Perceived Stress Score (PSS) for ambulatory setting, first proposed in [12] and subsequently used in [36].

Each item is scored on a scale of 1 to 6. Taken altogether, these five scores can be processed, as we show later in the paper, to represent a subjective measure of participant's perception/awareness of stress at that moment. Each EMA self-report is time-stamped, and is used as stress ground-truth, albeit noisy, in field validation of *cStress*.

Net Data Collected

In all three data sets — ***train***, ***test***, and ***field*** — the data for several participants were removed from analysis due to either missing signals, insufficient good quality data, and/or insufficient or erroneous self-reports or EMAs.

In the case of ***train***, out of 24 participants, three had missing RIP data, and were excluded from analysis. For the remaining 21 participants, the average/total number of person minutes was 73/1534. The number of participants used for self-reports-based validation was further filtered down to 19, because 2 had missing self-reports. The average/total number of self-reports in the lab was 13/247. In the case of ***test***, we use all 26 participants. For these participants, the average/total number of person minutes was 58/1501.

Finally, in the case of ***field*** data, the initial number of participants was 23, but 3 had insufficient good data or EMA. From the remaining 20 participants, the total number of usable self-report that had good quality data prior to self-reports was 1060. For predicting a self-report, all available and usable physiological data preceding the self-report were used.

DATA PROCESSING AND MODEL DEVELOPMENT

We now describe the details of data processing and modeling, including screening, cleaning, feature computation, and training of the machine learning model to produce *cStress*.

Data Processing

The first task in the processing pipeline is to unpack the packets, received wirelessly from the sensors, and to assign a time-stamp to each sample. To maintain time synchronization among all the data collected, whether they are embedded on the phone (e.g., GPS, self-report) or coming from wireless sensors, each data packet is timestamped as soon as it is received on the phone. This introduces complications in maintaining accurate timestamps, especially if some packets are lost, or time-stamping process gets delayed due to buffer delays. Such irregular warping of packet inter-arrival times can degrade quality of features computed in the later steps.

Time-stamp alignment/correction and Data Interpolation—To remedy this, we apply a dynamic-programming approach to correct the time-stamps. We first obtain the ideal timestamps, by noting the time between the first and last packet, and figuring in the sampling frequency of the sensor signal. These ideal timestamps act as scaffolding to which we optimally align the actual sample timestamps. The dynamic programming approach we use is similar to time-series alignment algorithms, e.g. the Dynamic Time Warping algorithm. It selects the alignment that minimizes the sum of squared differences between the ideal and actual timestamps.

The time-stamp correction process identifies any losses in the sensor data stream. If a small amount of data is lost, we interpolate the missing signal samples. We use cubic Hermite splines to interpolate the gaps, which is known to be appropriate for interpolating physiological measurements [32]. However, for gaps that are too wide, interpolation would fail to correctly reproduce the peaks and valleys. For example, if a gap spans several peaks,

simple spline interpolation would not reproduce the actual peaks, which might lead to wrong features. If the gap fits inside a peak or valley, interpolation can be a viable way to restore the peak or valley well enough to be detected by the peak/valley code. In our case, each packet consists of only 5 samples, and each packet corresponds to only 8% of an ECG or respiration cycle, hence 1 packet can easily be interpolated without significant loss in accuracy of locating peaks and valleys. We impute if 1 packet is lost in a burst, which reduces the data loss rate from 10% to less than 1.5% (i.e., most packet losses are 1 packet long).

Detecting and Excluding Physical Activity Confounds

Throughout the data analysis, we require accurate detection of time intervals with moderate-to-high physical activity, in order to account for physiological arousal due to physical activity rather than stress. We limit the application of the stress monitoring framework to low/no activity intervals, which we consider as a type of admission control. If majority of ten-second windows inside the minute are classified as moderate-to-high activity, we designate the entire minute as moderate-to-high physical activity and screen it out. To determine the presence of physical activity inside of each 10-second window, we use a simple threshold based activity detector using the 3-axis on-body accelerometer (placed on chest). The choice of a 10-second activity detection window and the threshold-based detector is based on the method proposed in [38].

Feature Computation

The next steps involve computing the feature representation of each one-minute window observation, which may be used as a training or test observation. The entire data stream is then split into one minute intervals, and for each minute we compute various time-domain, as well as frequency-domain, aggregate functions of these base features, e.g. statistical aggregates like mean, variance, standard deviation, etc. These one-minute aggregates serve as the final features of each one-minute observation that may be used as a training observation by the Support Vector Machines algorithm that trains *cStress*, or as a test observation to which we apply *cStress*.

ECG features—ECG signal processing includes three phases. First, we identify the acceptable portions of an ECG signal. A portion of an ECG signal is considered acceptable if it retains characteristic morphologies of standard ECG, i.e. contains identifiable P and T waves and QRS complexes. Otherwise, it is deemed as unacceptable. Figure 2 illustrates both types of ECG signals. Improper attachment of electrodes produces triangular shape signal. Flat signal appears when sensor is detached from the body completely.

Second, all R-peaks are detected automatically from the acceptable ECG portions using Pan and Tompkins's algorithm [34]. Accuracy of R-peak detection in lab (Minnesota lab) and field (Memphis Field) data is 98.6% and 97.01% (when compared with manual marking via visual inspection). The difference between two consecutive R peaks is the R-R interval or inter beat interval (IBI). One missed R peak will elongate the inter beat interval (IBI) by at least twice the mean IBI, or more, in case of multiple missed peaks. False detection of R-peak within two actual peaks will reduce the resultant IBI. Thus, one or more consecutively

missed R-peaks or spuriously detected non-existent R-peaks will result in invalid R-R intervals.

Algorithm 1

Algorithm for determining whether current candidate R-R interval is valid.

1 function **isRRintValid** $(RR_n, RR_{n-1}, RR_{n+1}, RR_k)$;
Inputs : RR_n: current candidate R-R interval; RR_{n-1}: previous candidate R-R interval; RR_{n+1}: next candidate R-R interval; $RR_k|k < n$: last valid R-R interval before RR_n
Output: RR_nvalid: is RR_n valid?
2 **if** $300ms \leq RR_{n-1} \leq 2000ms$ **then**
3 | **if** $|RR_n - RR_{n-1}| \leq CBD$ **then**
4 | | $RR_nvalid = true$;
5 | **else**
6 | | $RR_nvalid = false$;
7 | **end**
8 **else**
9 | **if** $|RR_n - RR_k| \leq CBD$ **then**
10 | | $RR_nvalid = true$;
11 | **else**
12 | | **if** $|RR_n - RR_{n-1}| \leq CBD$ ***and*** $|RR_n - RR_{n+1}| \leq CBD$ **then**
13 | | | $RR_nvalid = true$;
14 | | **else**
15 | | | $RR_nvalid = false$;
16 | | **end**
17 | **end**
18 **end**

We propose Algorithm 1 to improve the outliers detection method of [7] and illustrate it on ECG signal presented in Figure 2. We refer the reader to [7] for the definition of criterion

beat difference (CBD). Evaluation on real-life data shows that this new method detects outliers in R-R intervals with an accuracy of 99.04% in lab and 97.8% in field.

In the next step, we normalize the R-R intervals to remove any subject/session specific components from the distribution of the R-R intervals. A careful and robust normalization process calibrates the model to each person without the need for constructing a personalized model that would require extensive training before the model can be applied to any person not in the training set.

We normalize the R-R intervals, i.e. compute the z-score, using scale deviations winsorized mean and variance estimates, with the threshold parameter of 3 [50]. Winsorization limits the values of extreme outliers at the boundaries below and above the median of the data. This is an alternative to simply trimming the data and removing outliers, and aims to save of the information present in such extreme outliers. We ensure that the data used to compute the winsorized mean and variance only contain samples during low or no physical activity, so as to eliminate any bias that physical activity may impart to the mean of variance of a participant's physiological arousal.

Finally, we use the normalized R-R interval values, derived in the previous steps, to compute the R-R aggregated features for every one minute window. These features are listed in Table 1. Several of these aggregated features belong to the category of heart rate variability (HRV) features.

Respiration features—Breathing dynamics can be captured using respiratory inductive plethysmograph (RIP) which unobtrusively tracks the change of ribcage circumference during inhalation and exhalation of a breathing cycle. Respiration signals are largely affected by positioning of the chest band, physical movement, loosening of electrical connectors and slipping of the band from its expected location. As illustrated in Figure 3, we mark the signal acceptable as long as the signal follows sinusoidal pattern. Mere loosening of the chest band sometimes results in a low amplitude signal, but that is considered acceptable if it still retains the characteristic morphology of a respiration signal. Signal saturation to a point where variation is no longer detectable is considered unacceptable, which can be seen in a case where the sensor is detached from the body. We adopt a method proposed in [35] for determining acceptability of ECG and respiration signals.

After removing poor quality signal, we identify each cycle by locating peaks and valleys of accepted respiration signal. For that, we adopt the method used in [24]. First, the waveform is separated into breath cycles by identifying intercepts of a moving average curve with the inspiration and expiration branches of the waveform as shown in Figure 3. Peaks and valleys are defined, respectively, as the maximum and minimum between pairs of alternating inspiration and expiration intercepts. Second, if an inspiration or expiration amplitude is too small, ≤20% of the mean peak to valley amplitude, the associated pair of peak and valley is deleted. Empirically, we find that respiration duration varies from 0.9 sec (during heavy exercise, i.e., running) to 12.5 sec (i.e. conversation). When searching peaks and valleys, only time intervals from valley to valley that fall within the range of 0.9 sec to 12.5 sec are accepted. Otherwise, the peaks and valleys are ignored because they are considered not to be

real peaks or valleys but small bumps or noise. The respiration duration upper limit of 12.5 sec is adopted from [28], and is also supported by our own data, which include carefully marked episodes of conversation, loud reading, and public speaking. The lower limit of 0.9 sec is calculated using our data which include running, walking, sitting, standing and lying episodes. Furthermore, it is close to the value mentioned in [31].

For each cycle, we compute various base features that describe the characteristics of this respiration cycle. We use the following respiration features, described above and outlined in Table 2: inspiration duration, defined as the time between start and end of inspiration inside the cycle; expiration duration, defined as the time between the start and the end of the expiration portion; respiration duration, defined as the total time of the respiration cycle; ratio of inspiration to expiration duration; stretch, defined as the difference between the maximum (legitimate) amplitude and the minimum (legitimate) amplitude of the signal within a respiration cycle.

We also compute Respiratory Sinus Arrhythmia (RSA), which is another feature sometimes used in emotion classification (e.g. [48]). It is a multimodal feature derived from both ECG and respiration that describes the variability in RR intervals due to respiration. Inspiration and expiration are associated with changes in RR intervals that may be driven by a central brainstem circuit rather than being causally related to the expansion and contraction of the chest. RSA is computed by subtracting the shortest RR interval from the longest RR interval within each respiratory cycle.

Next, we normalize the features in a similar manner as we normalized the R-R intervals, using low activity winsorized mean and variance estimates. The scale deviations winsorization threshold parameter was set at 3, as earlier.

Finally, for each one-minute interval, we compute various statistical aggregates, listed in Table 2, of these normalized base respiration features. Additionally, we compute two other per-minute features: breath rate, simply defined as the number of respiration cycles per minute; and inspiration minute volume, which is the volume of air inhaled into the lungs in one minute, estimated by computing the area under the curve of the inspiration phases of the respiration cycles in the minute. As with ECG, these statistics are used as the final features of each one-minute window observation that will be used either to train or test *cStress*.

Model Training and Validation

Once the normalized aggregated features are computed, we proceed to the step of learning the parameters of *cStress*, which outputs a probability of stress-driven physiological activation for any one-minute window input. One last bit of processing before running the machine learning algorithm is scaling each input feature between 0 and 1. This is a standard step that can significantly improve the learning algorithm performance, particularly in kernel-based learning algorithms, which is what we are deploying.

The model is trained using the well-known Support Vector Machines (SVM) algorithm[3]. It can be described as an L_2-regularized loss minimization algorithm, with the loss function defined as a classic hinge-loss [42]. It is noted for its ability to learn high-capacity models,

owing to the so-called Kernel trick, whilst limiting potential overfitting, thanks to regularization of parameters. Thus, the algorithm is explicitly formulated to attempt to reduce both the bias and variance of the resulting model.

The user can control the bias-variance tradeoff with a choice of the Kernel function and soft-margin hyper-parameter C. In learning *cStress*, we used the popular *RBF* Kernel, which requires a value for the hyper-parameter γ. According to the usual interpretation of SVM, the learned model is a hyper-plane, defined in some high-dimensional function space, which optimaly separates the space of observations into two subspaces — one for each class of observations [8]. The hyper-plane is selected by the SVM algorithm to maximize the margin of separation.

By default, the model's output is unscaled, whose absolute value represents the canonical distance of the input observation from the separating hyper-plane. The distance is proportional to the confidence in its classification. Applying a standard technique, called Platt's scaling [37], transforms the output into a conditional probability. It works by passing the output through a specially fitted sigmoid function.

The learned model outputs the probability that the input window belongs to class 'stressed'. If we want a binary classification, we can choose a threshold, related to the classification bias, and any minute with probability of stress above this threshold is classified as stress. This threshold bias can be considered as another hyper-parameter that needs to be tuned.

Hyper-parameter tuning—The performance of the SVM algorithm is highly sensitive to the choice of C and RBF γ. To choose the best values of C and γ, we perform a basic grid-search. The performance is evaluated using cross-subject validation, whereby we test the minutes of each subject with a model trained using all other subjects' minutes. We chose the F1 score as the performance metric, due to its popularity in those classification applications, where one class is the primary class of interest, and where the function can be thought of as a retrieval or detection system. The F1 score can be defined as a harmonic average between recall and precision of inferring stress arousal. The threshold bias is also chosen on the basis of maximizing the F1 score in cross-subject validation.

Self-reported Stress Inference Model

cStress captures the instantaneous physiological response from stressors. Its validation in the lab setting demonstrates its efficacy in identifying these patterns of arousal. But, there is no analog of a lab protocol in the field setting, against which *cStress* can be validated. The gold standard for the field setting has traditionally been the self-reports the participant fills out periodically throughout the day. Each self-report occurs at a random time in the day. In these self-reports, the participant records, among many things, his/her feelings of stress, anger/frustration, happiness, cheerfulness, and sadness, using a 6-level scoring system. The reliability and validity of self-reports have been questioned, because they are subject to biases, fabrication, falsification, and lack of care in reporting. In addition, they rely on memory. Physiology often responds to stressors instantaneously, subsiding when the stressor

[3]We deploy the popular LIBSVM library [9]

has faded. However, the memory of stress may persist in the mind of the participant, which is what the self-report captures. Hence, there may be an arbitrary lag between the occurrence of a stressor and its capture on the self-report. For such reasons, self-reports have produced only a marginal correlation with biofluid assessments, such as stress hormone (cortisol), of 0.26–0.36 [5, 6]. Nevertheless, it is the most widely used measure for validation in the field setting.

To allow for arbitrary lag between the physiological response captured by *cStress* and the memory of a stress event captured in self-report, we've developed a Bayesian Network model of self-reported stress that is similar to that proposed in [36].

Figure 4 illustrates the Bayesian Network model, which describes how self-reported stress values change in the course of the day at one-minute intervals. The model formalizes the recursive relationship between the (estimated) self-reported stress at any one minute of the day and the previous minute's (estimated) self-reported-stress, as well as the previous minute's physiological stress arousal (obtained from *cStress*). There are three variables in the model: S_i, S_{i-1}, and Z_{i-1}. All three variables are binary, valued as 1 ('stressed') or 0 ('not stressed'). S_i and S_{i-1} represent the self-reported stress for minute i (current minute) and minute i 1, respectively, and Z_{i-1} can be defined as physiological stress arousal at minute i −1. The network's connections are based on the notion that perception of stress for minute i, S_i, depends on perception of stress in the previous minute, S_{i-1}, and on whether there was physiological stress arousal in the previous minute, Z_{i-1}.

The probability distributions used in the model are the following: $p(S_i|S_{i-1}, Z_{i-1})$, $p(S_{i-1})$, and $p(Z_{i-1})$, where $p(S_i|S_{i-1}, Z_{i-1})$ is given by the conditional probability table (CPT) in Figure 4. Note that we simplified the parameterization of $p(S_i|S_{i-1}, Z_{i-1})$ by setting the probability of self-reported stress for minute i to 1 if there was also self-reported stress for minute i–1 and physiological stress activation for minute i–1, as measured by *cStress*. Conversely, if there was no detection of physiological stress activation for minute i–1, nor self-reporting of stress for minute i–1, then the probability of self-reported stress for minute i is 0. This simplification is both logical, and leads to a model of just two parameters, α and β.

The prior probability $p(Z_{i-1})$ is produced directly by *cStress*, whereas the marginal $p(S_{i-1})$, or in general $p(S_i)$ for any i, can be computed by marginalizing it from the joint distribution $p(S_i, S_{i-1}, Z_{i-1})$:

$$\begin{aligned} &p(S_i{=}1){=} \\ &p(S_i{=}1|S_{i-1}{=}0, Z_{i-1}{=}0)p(S_{i-1}{=}0)p(Z_{i-1}{=}0){+} \\ &p(S_i{=}1|S_{i-1}{=}1, Z_{i-1}{=}0)p(S_{i-1}{=}1)p(Z_{i-1}{=}0){+} \\ &p(S_i{=}1|S_{i-1}{=}0, Z_{i-1}{=}1)p(S_{i-1}{=}0)p(Z_{i-1}{=}1){+} \\ &p(S_i{=}1|S_{i-1}{=}1, Z_{i-1}{=}1)p(S_{i-1}{=}1)p(Z_{i-1}{=}1) \end{aligned}$$

The above equation can be simplified, using the CPT in Figure 4. Referring to $p(S_i = 1)$ simply as y_i and $p(Z_i = 1)$ as x_i, we have the following:

$$p(S_i=1)=y_i=\alpha y_{i-1}(1-x_{i-1})+\beta(1-y_{i-1})x_{i-1}+y_{i-1}x_{i-1} \quad (1)$$

We initialize this recurrence chain with the first self-report of the day, S_0.

$$p(S_1=1)=y_1=\begin{cases} \alpha(1-x_0)+x_0, & \text{if } S_0=1 \\ \beta x_0, & \text{otherwise.} \end{cases} \quad (2)$$

Equations (1) and (2) can be used to compute all the marginal probabilities of self-reported stress. Based on these marginals, we can classify each hypothetical self-report for every minute of the day as 'stressed' or 'not stressed'. The learning of α and β is performed using the available self-reports for each participant. Thus, each participant has his/her own unique α and β, learned using only that participant's field data and self-reports. Furthermore, for each participant there is just one α and β spanning all of his/her field study days.

To learn the model, we use a grid-search for α and β that maximize the F1 score of classifying all actual self-reports into either class 'stressed' or 'not stressed'. To compute the F1 score, we need the probabilities $p(S_i = 1)$, computed using equations (1) and (2), and the ground truth labels for S_i, which can be computed from the EMA self-report scores, by quantizing them into binary 'stressed'/'not stressed' labels.

To binarize self-report scores, we first average across all 5 stress items, reverse coding the two positive items (i.e., "happy" and "cheerful"). Next, we compute the mean of this quantity (i.e., *score*), for each participant, and use this mean as a threshold. For every score above the mean, we classify the self-report as "stressed", and "not stressed", otherwise.

EXPERIMENTAL RESULTS

In this section, we present the results of validation experiments on all three data sets. For the two datasets that have lab ground-truth labels — ***train*** and ***test*** — we perform standard classification experiments and report standard classification performance measures. However, the experimental designs for the two cohorts differ somewhat. For ***train***, we performed cross-subject validation, allowing us to fine tune the hyperparameters, and learn the final *cStress* model. This model was then used to perform out-of-sample validation on ***test***.

To evaluate the performance, we use standard performance measures: F1 score, which is also the measure based on which we tune the hyper-parameters C and γ; area under ROC curve (AUC); Accuracy, comprised of the Percent Correct, True Positive Rate, and False Positive Rate; and Cohen's Kappa. To understand and compare the predictive powers of different types of features, we repeated the experiments for the following categories/sets of features: entire set of 37 ECG and RIP features, just the ECG features, just the HRV features, and just the RIP features. Table 3 lists the values of all these performance measures for cross-subject validation on ***train***, for all four categories of features. The table also lists the optimal hyper-

parameters: C, γ, and *bias*. For additional reference, Table 4 contains the confusion matrix of the cross-subject validation tests using the optimal hyper-parameters and all features.

In the case of ***test***, we classified the minutes of each participant separately, producing the list of performance measures for each participant. We used *cStress* trained on all features using the corresponding C and γ. We obtain a median accuracy of 95.3%, AUC of 0.98, Kappa of 0.87, MCC of 0.88, and F1 score of 0.9. The sets of performance measures are plotted in box plots in Figure 5. Additionally, we compiled a confusion matrix, seen in Table 5, made up of the combination of confusion matrices of all participants.

For the final set of experiments, we fit the Bayesian Network model of self-reports to self-reported stress scores in the lab and in the field. The model is fitted for each participant separately. The objective of these experiments is to validate *cStress* with instantaneous self-reports. This experiment is performed on ***train*** *and* ***field***, both of which contain self-reported stress scores. Table 6 contains the median values of performance measures, F1, Accuracy, and AUC, across all participants, for both data sets. As the table shows, using a simple two-parameter model, we are able to relatively accurately infer self-reports, especially given the limitations of self-reports, as discussed earlier. Additionally, in Figures 6 and 7, we present the accuracy values for each participant separately, for both data sets ***train*** and ***field***, respectively.

Finally, we present a ranking of features in terms of contribution to the model performance. We employ a variant of Multiple Kernel Learning (MKL), called *simpleMKL* [41], with a separate Kernel for each feature, to rank features based on their associated Kernel weight coefficients in the final *simpleMKL* model. As we did in the previous experiments, we use grid-search to fine-tune the hyper-parameters of the *sim-pleMKL* algorithm to maximize the cross-subject validation F1 score. The best result we obtain has F1=0.81, AUC=0.96, and Accuracy=0.93, which is similar to what was achieved with the SVM model using all features, reported earlier in Table 3. After this, the best hyper-parameter values are used to learn the final *simpleMKL* model, using the entire training data, from which we can extract the learned Kernel weight coefficients. Figure 8 shows a bar plot of these weight coefficients for all features, sorted from biggest to smallest. The high level of sparsity is due to the ℓ_1-norm regularization of Kernel coefficients performed by *simpleMKL*.

RELATED WORKS

There is a rich body of related works on stress assessment. Most of these have been done in a controlled lab setting or supervised real-life setting. In the first case, the research efforts are focused on discovery and analysis of effective indicators of stress [49, 46, 19, 10, 51, 39]. These works contain highly useful findings and analyses, in particular, effective features. One example of this is the widespread adoption of heart rate variability (HRV) features [25, 29, 49]. We evaluated the utility of HRV features in measuring stress and find that on the lab training data, using HRV features alone produces an F1 score of 0.56 as compared with 0.81 when all the features of *cStress* model are used (see Table 3).

Several works [23, 25, 29] report on experiments in a real-world scenario. The number of situations and activities, however, are usually limited. For instance, in [25], stress was inferred only while the participants were on the computer. Similarly, in [29], the participants participated in only one type of stressful situation – verbal examination – and the non-stressful period took place in a controlled rest setting. Two of these papers used heart rate variability features as a measure of stress arousal [25, 29].

The last category of papers are the papers that discuss stress monitoring in the wild [22, 36, 2]; our work also belongs to this class. In [22], the authors present a review of other papers on stress inference in the wild, as well as discuss their own efforts in this direction. They propose a feature called additional heart rate (AHR), which has been found to be predictive of stress [30]. However, the paper mentions that the authors did not analyze the accuracy, and seemingly performed only limited validation on field data.

The closest to our work is [36] that used a similar lab setup for data collection and for training their model. They experienced data quality issues and excluded majority of data used for training and testing. They used only 28 minutes (= 600/21) of data per participant in comparison to 73 minutes per participant for our case, for the same protocol where each participant spent 103 minutes in the lab (see Figure 3 in [36]). Despite careful data exclusion, [36] reported a recall rate of 88% and a false positive rate of 8% on training data set. The field data had similar issues — 66% of the field data was excluded, leaving only 16 hours of data per participant. In contrast, we use 50+ hours of good quality data per participant, which are all independent from the training set [38].

Most importantly, when validating against self-report, they only compared against an overall average stress level (aggregated over 2 days) for each participant. As a result, they only had one data point (pair of model output and self-report) for each participant. This does not indicate whether the model can predict the instantaneous self-reports and hence limits the utility of the model for producing a continuous measure of stress. In contrast, we use 53 self-reported data per participant and show that *cStress* is capable of inferring each self-reported stress. To the best of our knowledge, ours is the first work to propose a stress model that has been validated on independent data sets in both lab and the field and is able to predict each instantaneous self-report collected in the field.

DISCUSSION, CONCLUSIONS AND FUTURE WORK

Our proposed *cStress* model obtains good accuracy on independent data sets in both lab and field and constitutes a significant step towards a gold standard for continuous stress assessment from wearable wireless sensors.

This work, however, has several limitations and significant potential for future works. First, for *cStress* to truly become a gold standard, it needs further improvements in accuracy and reproduction on other independent data sets. Several approaches could be adopted to improve the accuracy such as more convenient data collection methods (e.g., obtaining inter-beat intervals from smartwatches instead of ECG electrodes), better handling of physical

activity confounds so fewer data segments are filtered out, personalization of the model to the context, among several others.

Second, to become societally useful, its clinical utility in the management of stress needs to be established. For example, sensor-triggered just-in-time mobile interventions for stress management could be developed and evaluated among those suffering from migraine, stress disorders, or those abstaining from addictive behaviors (e.g., smoking). Third, effective visualizations could be developed that permit users to visualize their stress patterns on mobile devices and gain insights into contexts that may increase or decrease their daily stress.

Fourth, when stress assessment is combined with other data such as geoexposures (from GPS), visual exposures (from smart eyeglasses), social interactions (from microphones), light and sound exposures (from smartwatch sensors), and digital trails (from social media, emails, calendars, etc.), stress predictors could be discovered for better management of daily stress. Finally, stress may be a socially private information for some. Hence, it raises new privacy management issues for mobile sensor data.

Acknowledgments

We thank Rummana Bari, Monowar Hossain, Mahbubur Rahman, Moushumi Sharmin, and Hillol Sarker from the University of Memphis. Rummana made substantial contributions to the data screening and cleaning methods, and others contributed to the data analysis methods and manuscript review. We also thank the study coordinators at both the University of Minnesota and the University of Memphis. The authors acknowledge support by the National Science Foundation under award numbers CNS-1212901 and IIS-1231754, and by the National Institutes of Health under grants R01DA035502 (NIDA) through funds provided by the trans-NIH OppNet initiative and U54EB020404 (NIBIB) through funds provided by the trans-NIH Big Data-to-Knowledge (BD2K) initiative.

References

1. Zephyr Bioharness. [Accessed: September 2013] http://www.zephyr-technology.com/bioharness-bt
2. Adams P, Rabbi M, Rahman T, Matthews M, Voida A, Gay G, Choudhury T, Voida S. Towards personal stress informatics: Comparing minimally invasive techniques for measuring daily stress in the wild. Pervasive Health. 2014:72–79.
3. al'Absi, M. Stress and addiction: Biological and psychological mechanisms. Academic Press/Elsevier; 2007.
4. al'Absi M, Arnett D. Adrenocortical responses to psychological stress and risk for hypertension. Biomedecine & Pharmacotherapy. 2000; 54(5):234–244.
5. al'Absi M, Bongard S, Buchanan T, Pincomb G, Lovallo JLW. Cardiovascular and neuroendocrine adjustment to public speaking and mental arithmetic stressors. Psychophysiology. 1997; 34:266–75. [PubMed: 9175441]
6. al'Absi M, Hatsukami D, Davis G, Wittmers L. Prospective examination of effects of smoking abstinence on cortisol and withdrawal symptoms as predictors of early smoking relapse. Drug Alcohol Dependence. 2004; 73(3):267–78. [PubMed: 15036549]
7. Berntson G, Quigley K, Jang J, Boysen S. An approach to artifact identification: Application to heart period data. Psychophysiology. 1990; 27(5):586–598. [PubMed: 2274622]
8. Boser, BE., Guyon, IM., Vapnik, VN. A training algorithm for optimal margin classifiers. Fifth Annual Workshop on Computational Learning Theory; 1992. p. 144-152.
9. Chang C-C, Lin C-J. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology. 2011; 2:27:1–27:27. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

10. Choi J, Ahmed B, Gutierrez-Osuna R. Development and evaluation of an ambulatory stress monitor based on wearable sensors. IEEE Transactions on Information Technology in Biomedicine. 2012; 16(2):279–286. [PubMed: 21965215]
11. Chrousos G, Gold P. The concepts of stress and stress system disorders: overview of physical and behavioral homeostasis. JAMA. 1992; 267(9):1244. [PubMed: 1538563]
12. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. Journal of Health and Social Behavior. 1983; 24(4):385–396. [PubMed: 6668417]
13. Enoch M. Pharmacogenomics of alcohol response and addiction. American Journal of Pharmacogenomics. 2003; 3(4):217–232. [PubMed: 12930156]
14. Enoch M. Genetic and environmental influences on the development of alcoholism. Ann NY Acad Sci. 2007; 1094:193–201.
15. Ertin E, Stohs N, Kumar S, Raij A, al'Absi M, Kwon Mitra TS, Shah S, Jeong J. AutoSense: Unobtrusively Wearable Sensor Suite for Inferencing of Onset, Causality, and Consequences of Stress in the Field. ACM SenSys. 2011
16. Healey J, Nachman L, Subramanian S, Shahabdeen J, Morris M. Out of the lab and into the fray: Towards modeling emotion in everyday life. Pervasive Computing. 2010; 6030:156–173.
17. Healey JA, Picard RW. Detecting stress during real-world driving tasks using physiological sensors. IEEE Transactions on Intelligent Transportation Systems. 2005; 6(2):156–166.
18. Henry J. Stress, neuroendocrine patterns, and emotional response. Stressors and the adjustment disorders. 1990:477–496.
19. Hong J-H, Ramos J, Dey AK. Understanding physiological responses to stressors during physical activity. ACM UbiComp. 2012:270–279.
20. Kreibig SD. Autonomic nervous system activity in emotion: A review. Biological Psychology. 2010; 84(3):394–421. [PubMed: 20371374]
21. Kreibig SD, Wilhelm FH, Roth WT, Gross JJ. Cardiovascular, electrodermal, and respiratory response patterns to fear- and sadness-inducing films. Psychophysiology. 2007; 44(5):787–806. [PubMed: 17598878]
22. Kusserow M, Amft O, Troster G. Monitoring stress arousal in the wild. IEEE Pervasive Computing. 2013; 12(2):28–37.
23. Lu H, Frauendorfer D, Rabbi M, Mast MS, Chittaranjan GT, Campbell AT, Gatica-Perez D, Choudhury T. Stresssense: Detecting stress in unconstrained acoustic environments using smartphones. ACM UbiComp. 2012:351–360.
24. Lu W, Nystrom MM, Parikh PJ, Fooshee DR, Hubenschmidt JP, Bradley JD, Low DA. A semi-automatic method for peak and valley detection in free-breathing respiratory waveforms. Medical Physics. 2006; 33(10):3634–3636. [PubMed: 17089828]
25. Mark G, Wang Y, Niiya M. Stress and multitasking in everyday college life: An empirical study of online activity. ACM CHI. 2014:41–50.
26. McEwen B. Protection and damage from acute and chronic stress. Ann NY Acad Sci. 2004; 1032:1–7. [PubMed: 15677391]
27. McEwen B, Stellar E. Stress and the individual: mechanisms leading to disease. Archives of Internal Medicine. 1993; 153(18):2093. [PubMed: 8379800]
28. Mcfarland DH. Respiratory Markers of Conversational Interaction. Journal of Speeh, Language and Hearing Research. 2001; 44:128–143.
29. Melillo, P., Formisano, C., Bracale, U., Pecchia, L. World Congress on Medical Physics and Biomedical Engineering. Springer; Berlin Heidelberg: 2013. Classification tree for real-life stress detection using linear heart rate variability analysis. case study: students under stress due to university examination; p. 477-480.
30. Myrtek, M. Heart and emotion: Ambulatory monitoring studies in everyday life. Hogrefe & Huber Publishers; 2004.
31. Neder J, Dal Corso S, Malaguti C, Reis S, De Fuccio M, Schmidt H, Fuld J, Nery L. The pattern and timing of breathing during incremental exercise: a normative study. European Respiratory Journal. 2003; 21(3):530–538. [PubMed: 12662013]

32. Nielsen P, Grice IL, Smaill B, Hunter P. Mathematical model of geometry and fibrous structure of the heart. American Journal of Physiology-Heart and Circulatory Physiology. 1991; 260(4):H1365–H1378.
33. Nunnally, JC. Psychometric Theory. 2. McGraw-Hill; New York: 1978.
34. Pan J, Tompkins W. A real-time qrs detection algorithm. IEEE Transactions on Biomedical Engineering. 1985; 32(3):230–236. [PubMed: 3997178]
35. Plarre K, Raij A, Guha S, al'Absi M, Ertin E, Kumar S. Automated detection of sensor detachments for physiological sensing in the wild. Wireless Health. 2010; 2010:216–217.
36. Plarre K, Raij A, Hossain S, Ali A, Nakajima M, al'Absi M, Ertin E, Kamarck T, Kumar S, Scott M, Siewiorek D, Smailagic A, Wittmers L. Continuous inference of psychological stress from sensory measurements collected in the natural environment. ACM IPSN. 2011:97–108.
37. Platt, JC. Advances in Large Margin Classifiers. MIT Press; 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods; p. 61-74.
38. Rahman M, Bari R, Ali A, Sharmin M, Raij A, Hovsepian K, Hossain S, Ertin E, Kennedy A, Epstein D, et al. Are we there yet?: feasibility of continuous stress assessment via wireless physiological sensors. ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (BCB). 2014:479–488.
39. Rahman T, Zhang M, Voida S, Choudhury T. Towards accurate non-intrusive recollection of stress levels using mobile sensing and contextual recall. Pervasive Health. 2014:166–169.
40. Rainville P, Bechara A, Naqvi N, Damasio AR. Basic emotions are associated with distinct patterns of cardiorespiratory activity. International journal of psychophysiology. 2006; 61(1):5–18. [PubMed: 16439033]
41. Rakotomamonjy A, Bach FR, Canu S, Grandvalet Y, Simplemkl. Journal of Machine Learning Research. 2008; 9:2491–2521.
42. Rifkin, R. PhD thesis. MIT; 2002. Everything Old is New Again: A Fresh Look at Historical Approaches in Machine Learning.
43. Rosmond R, Björntorp P. Endocrine and metabolic aberrations in men with abdominal obesity in relation to anxio-depressive infirmity. Metabolism. 1998; 47(10):1187–1193. [PubMed: 9781619]
44. Sapolsky, RM. Why zebras don't get ulcers: The acclaimed guide to stress, stress-related diseases, and coping-now revised and updated. Macmillan; 2004.
45. Sarker, H., Sharmin, M., Ali, AA., Rahman, MM., Bari, R., Hossain, SM., Kumar, S. Assessing the availability of users to engage in just-in-time intervention in the natural environment. Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp '14, ACM; New York, NY, USA. 2014. p. 909-920.
46. Sharma, N., Gedeon, T. Advances in Knowledge Discovery and Data Mining. Springer; Berlin Heidelberg: 2013. Computational models of stress in reading using physiological and physical sensor data; p. 111-122.
47. Sharmin M, Raij A, Epstein D, Nahum-Shani I, Beck G, Vhaduri S, Preston K, Kumar S. Visualization of time-series sensor data to inform the design of just-in-time adaptive stress interventions. ACM UbiComp. 2015
48. Stephens C, Christie I, Friedman B. Autonomic specificity of basic emotions: Evidence from pattern classification and cluster analysis. Biological psychology. 2010; 84(3):463–473. [PubMed: 20338217]
49. Sun F-T, Kuo C, Cheng H-T, Buthpitiya S, Collins P, Griss M. Activity-aware mental stress detection using physiological sensors. Mobile Computing, Applications, and Services. 2012; 76:211–230.
50. Wu, M. PhD thesis. Michigan State University; 2006. Trimmed and Winsorized Eestimators.
51. Zhai J, Barreto A. Stress detection in computer users based on digital signal processing of noninvasive physiological variables. IEEE EMBS. 2006:1355–1358.

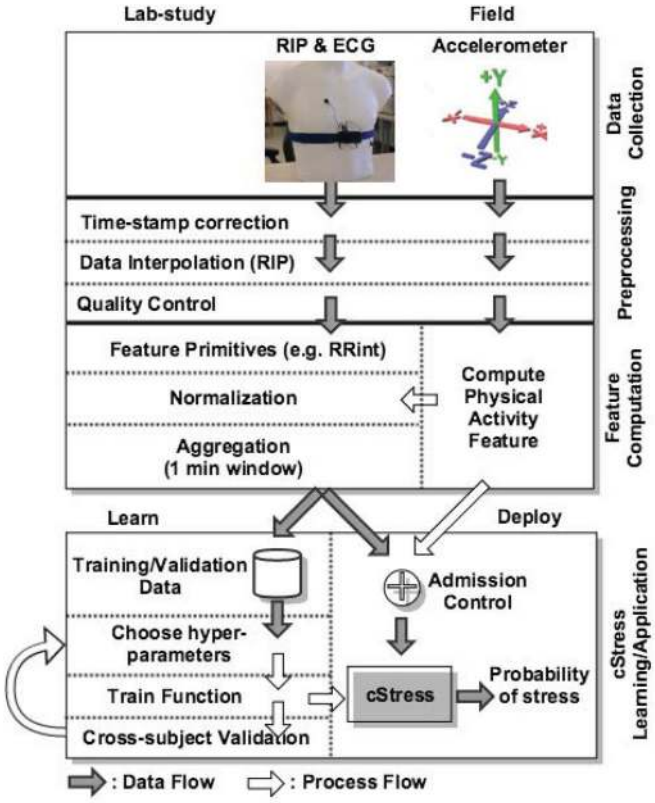


Figure 1.
Overview of the data processing and machine learning steps.

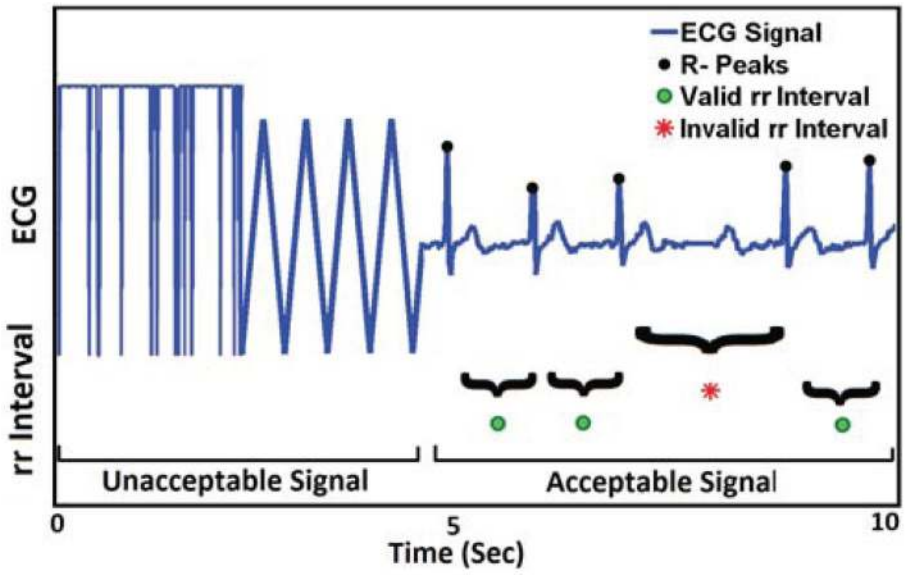


Figure 2.

The portion of signal that holds ECG property marked as acceptable. The triangular shape and saturated at top is labeled unacceptable. Increased R-R interval due to missed R peaks are detected as invalid by the algorithm and marked with red dot.

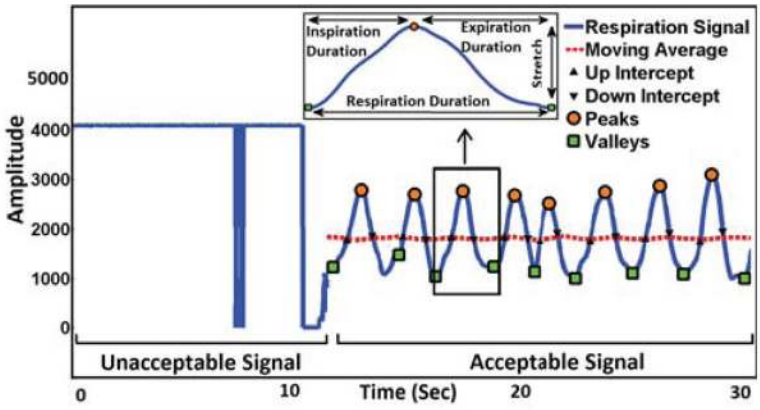
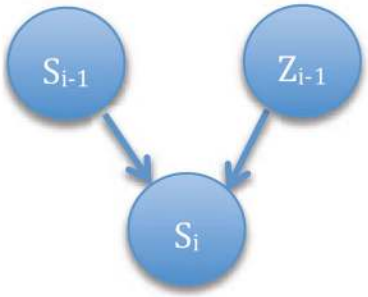


Figure 3.

Illustration of acceptable/unacceptable RIP signals, and computaton of base RIP features. The portion of signal that holds respiration signal property, looks like quasi-sinusoidal is marked as acceptable. Saturated at top is labeled unacceptable.



		S_i	
S_{i-1}	Z_{i-1}	**1**	**0**
1	**1**	1	0
1	**0**	α	1-α
0	**1**	β	1-β
0	**0**	0	1

Figure 4.

On the left, the Bayesian Network that explains the causal relationship between the self-reported stress for minute $i-1$, the physiological stress arousal for minute $i-1$ and the self-reported stress for minute i. On the right, the conditional probability table for S_i.

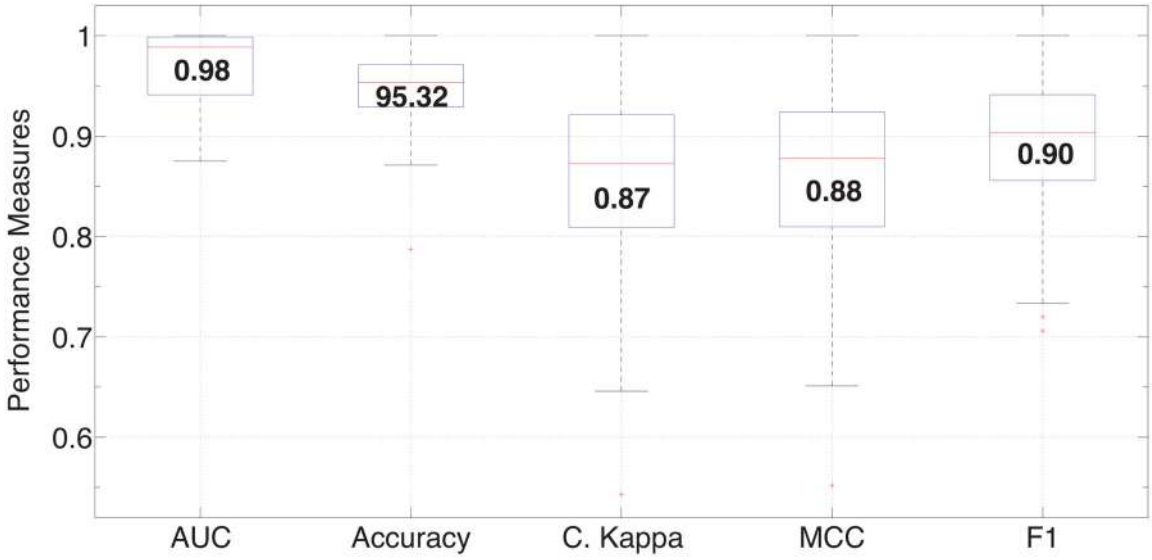


Figure 5.

Box plots of AUC, Accuracy, Cohen's Kappa, Matthew's Correlation Coeffiecient (MCC), F1 score for all 26 lab-study participants in the *test* cohort. Median values are displayed inside each box.

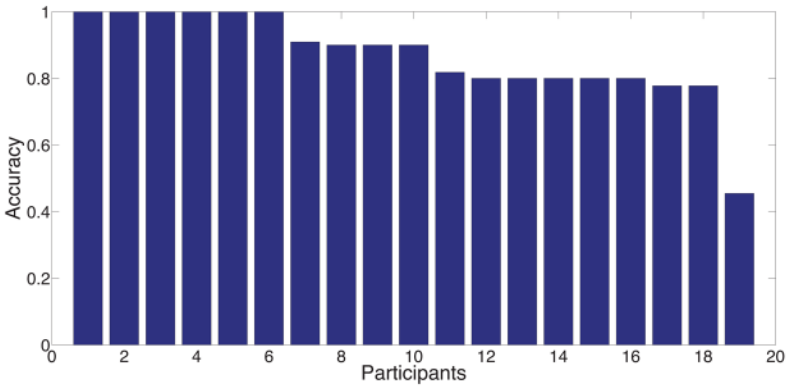


Figure 6.
Self-reported stress inference performance measures for each participant in the *train* cohort.

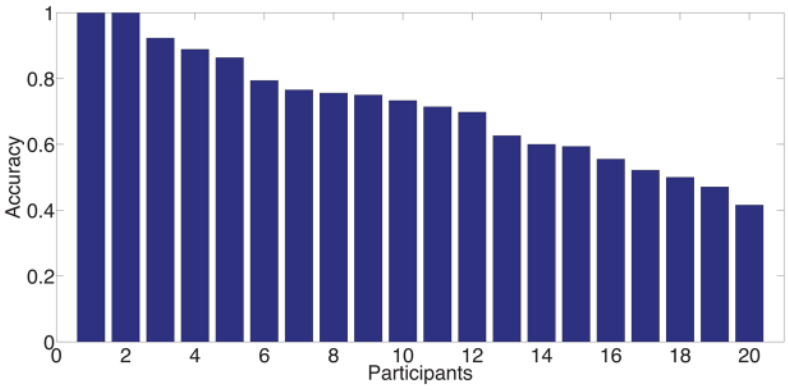


Figure 7.
Self-reported stress inference performance measures for each participant in the *field* cohort.

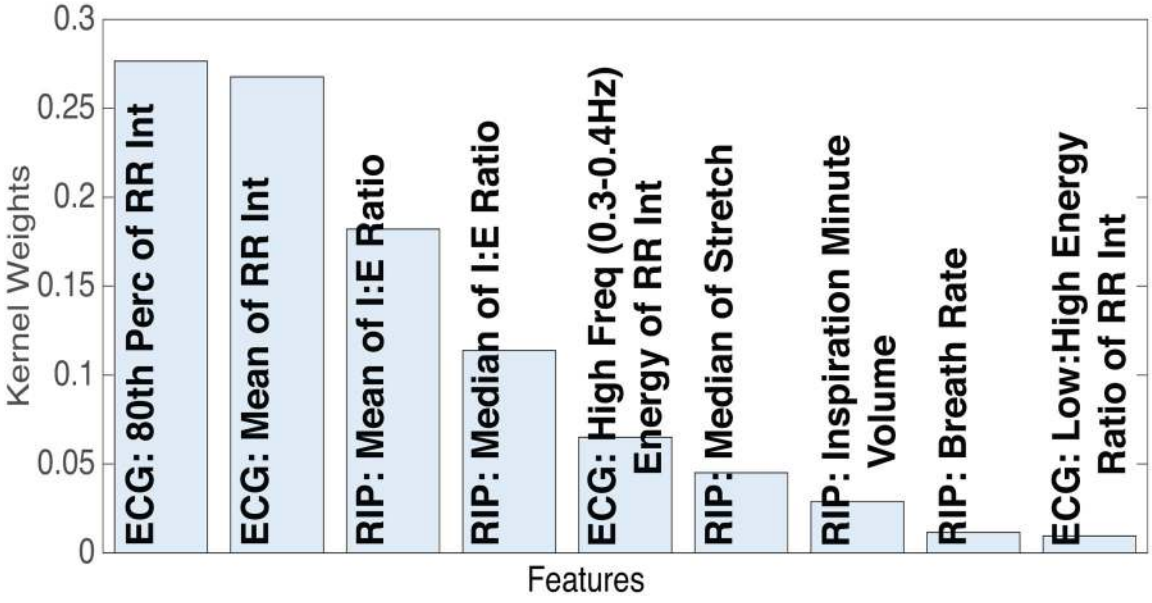


Figure 8.

Plot of MKL Kernel weight coefficients, sorted from biggest to smallest. Each weight corresponds to a separate feature, based on which the corresponding Kernel was computed. For the sake of space, we cut out all features below the weight coefficient of 0.01.

Table 1

All aggregated ECG features, computed using the processed (filtered and normalized) R-R intervals. The table mentions which of the ECG features are HRV features.

HRV	variance, quartile deviation, low frequency energy (0.1–0.2Hz), medium frequency energy (0.2–0.3Hz), high frequency energy (0.3–0.4Hz), low:high frequency energy ratio
non-HRV	mean, median, 80th percentile, 20th percentile, heart-rate

Table 2

All of the base and aggregated RIP features, which are computed by our system.

Base Features	**Aggregations**
inspiration duration, expiration duration, respiration duration, I:E duration ratio, stretch, respiratory sinus arrhythmia (RSA)[1]	mean, median, 80th percentile, quartile deviation
breath-rate[2], inspiration minute volume[2]	

[1]: RSA is a hybrid feature that uses both RIP and ECG signals.

[2]: The aggregated features breath rate and inspiration minute volume are computed without any other base features, but rather using just the number of respiration cycles in a minute.

Table 3

Cross-subject validation performance metrics for dataset *train*

			Accuracy				Optimal hyper-parameters		
Feature Set	**F1**	**AUC**	**Hit-rate**	**TPR**	**FPR**	**C. Kappa**	***C***	***γ***	***bias***
All	0.81	0.96	0.93	0.84	0.05	0.77	90.5097	0.000345267	0.339329
ECG	0.78	0.95	0.92	0.72	0.05	0.73	2	0.00552427	0.340407
HRV	0.56	0.78	0.84	0.55	0.1	0.46	724.077	0.0220971	0.250926
RIP	0.75	0.93	0.90	0.83	0.09	0.69	1448.15	0.000488281	0.308312

Table 4

Cross-subject validation confusion matrix for for dataset *train*

		Classified by Model		
		Stressed	**Not stressed**	**Total**
Actual	Stressed	236 (84%)	46 (16%)	282
	Not stressed	61 (5%)	1191 (95%)	1252
	Total	291	1237	1534

Table 5

Test confusion matrix for dataset *test*, made by combining the confusion matrices for all test participants.

		Classified By Model		
		Stressed	**Not stressed**	**Total**
Actual	Stressed	351 (89%)	45 (11%)	396
	Not stressed	56 (5%)	1149 (95%)	1205
	Total	407	1194	1501

Table 6

Median self-reported stress inference results, across all participants, for *train* and *field* cohorts.

	train	**field**
Median F1	0.75	0.71
Median AUC	0.85	0.60
Median Accuracy	0.9	0.72