micro-Stress EMA: A Passive Sensing Framework for Detecting in-the-wild Stress in Pregnant Mothers

ZACHARY D. KING, Northwestern University, United States
JUDITH MOSKOWITZ, Northwestern University, United States
BEGUM EGILMEZ, Northwestern University, United States
SHIBO ZHANG, Northwestern University, United States
LIDA ZHANG, Northwestern University, United States
MICHAEL BASS, Northwestern University, United States
JOHN ROGERS, Northwestern University, United States
ROOZBEH GHAFFARI, Northwestern University, United States
LAURIE WAKSCHLAG, Northwestern University, United States
NABIL ALSHURAFA, Northwestern University, United States

High levels of stress during pregnancy increase the chances of having a premature or low-birthweight baby. Perceived self-reported stress does not often capture or align with the physiological and behavioral response. But what if there was a self-report measure that could better capture the physiological response? Current perceived stress self-report assessments require users to answer multi-item scales at different time points of the day. Reducing it to one question, using microinteraction-based ecological momentary assessment (micro-EMA, collecting a single *in situ* self-report to assess behaviors) allows us to identify smaller or more subtle changes in physiology. It also allows for more frequent responses to capture perceived stress while at the same time reducing burden on the participant. We propose a framework for selecting the optimal micro-EMA that combines unbiased feature selection and unsupervised Agglomerative clustering. We test our framework in 18 women performing 16 activities in-lab wearing a Biostamp, a NeuLog, and a Polar chest strap. We validated our results in 17 pregnant women in real-world settings. Our framework shows that the question "How worried were you?" results in the highest accuracy when using a physiological model. Our results provide further in-depth exposure to the challenges of evaluating stress models in real-world situations.

ACM Reference Format:
Zachary D. King, Judith Moskowitz, Begum Egilmez, Shibo Zhang, Lida Zhang, Michael Bass, John Rogers, Roozbeh Ghaffari, Laurie Wakschlag, and Nabil Alshurafa. 2019. micro-Stress EMA: A Passive Sensing Framework for Detecting in-the-wild Stress in Pregnant Mothers. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 3, Article 91 (September 2019), 22 pages. https://doi.org/10.1145/3351249

Authors' addresses: Zachary D. King, Northwestern University, Evanston, IL, United States; Judith Moskowitz, Northwestern University, Chicago, IL, United States; Begum Egilmez, Northwestern University, Evanston, IL, United States; Shibo Zhang, Northwestern University, Evanston, IL, United States; Lida Zhang, Northwestern University, Evanston, IL, United States; Michael Bass, Northwestern University, Chicago, IL, United States; John Rogers, Northwestern University, Evanston, IL, United States; Roozbeh Ghaffari, Northwestern University, Evanston, IL, United States; Laurie Wakschlag, Northwestern University, Chicago, IL, United States; Nabil Alshurafa, Northwestern University, 680 N Lake Shore Dr, Chicago, IL, 60611, United States nabil@northwestern.edu.

2474-9567/2019/9-ART91 $15.00
https://doi.org/10.1145/3351249

1 INTRODUCTION

While immediate short-term mental and physical stress can be beneficial to health, prolonged elevated stress levels negatively affect the body's respiratory, cardiovascular, digestive, muscular, reproductive, nervous, and immune systems [18, 50, 54]. Prolonged stress can also contribute to poor health behaviors such as overeating [45], smoking [34], alcohol [42] and substance abuse [46]. These behaviors, in turn, are associated with conditions such as hypertension, heart disease, and depression [8]. Indeed, stress exposure in utero has been linked to deleterious outcomes for the child including impaired motor development, lower mental development, heightened behavioral disinhibition, and associated clinical patterns such as attention-deficit/hyperactivity disorder and disruptive behavior[5, 9, 14, 37]. There is a critical need to detect such exposures to stress in real-world situations in order to deliver timely stress-reduction interventions to improve maternal well-being and offspring neurodevelopment.

Stress is a complex concept that is often measured through self-report questionnaires. However, self-report measures are prone to several forms of recall bias and can be burdensome [20]. Further, for most individuals, perceived stress fluctuates over time and in response to different psychosocial stressors. An experience sampling approach, ecological momentary assessment (EMA), mitigates recall bias by repeatedly prompting participants to report their behavior, affect, and experience in close proximity to the event and in their usual environment. Participants are prompted by brief smartphone surveys multiple times (typically 6-10) each day. Although EMA may be considered a gold-standard approach for *in situ* data collection, participant burden remains a significant limitation [20].

Use of microinstruction EMA (micro-EMA) through smartwatches or smartphones (using a quick glance and tap method) has been shown to be less burdensome when prompting users to respond to questions and has shown increased participant response compared to timed multi-question EMAs [44]. Most multi-question EMAs do not allow for studying stress at a granular level, in the moment, multiple times per day, or over long durations due to the burden of multi-item scales [31]. A micro-EMA provides the means for participants to answer a single question more frequently throughout the day with less burden compared with the typical multi-question response required for a stress scale. A framework is needed that can identify the optimal single question to be used to establish a measurable ground truth that closely aligns with physiologic stress and correlates best with self-report. With reasonable correlation between physiology and self-report, health psychologists and others interested in the health effects of stress are more likely to begin to adopt passive sensing of stress.

With the rise of wearable sensors, there is promise in measuring physiologic stress using passively detectable, objective, and unbiased methods [21]. Inducing stress in participants in a controlled laboratory setting allows the use of machine learning algorithms to build stress-detection models that map physiologic signals (e.g., galvanic skin response [GSR], heart rate variability, and interbeat [RR] interval [IBI] signals) to stress in real time. There have been multiple studies that propose a machine learning model to predict stress using a range of different wearable sensors [7, 28, 43]. The in-lab portions of these studies involve a participant performing a structured activity and then reporting their stress before and after the activity. However, given limited validated stressors in research, the majority of these in-lab studies have little variation in the activities they define as stressful. For instance, the Trier social stress test (TSST) [33] is a proven and commonly used stress-induction method. The TSST is known to produce a consistent hypothalamic-pituitary-adrenal (HPA) axis response in humans, and dysfunction of this axis is associated with physical and mental health disorders [41]. Other stress induction techniques, such as the Stroop test [12] and the sing-a-song stress test [4], tend to produce variable or no HPA axis response in humans. It was recently determined that the key elements of stressors that produce HPA axis response include the combination of social-evaluative threat and lack of controllability [16]. It is also well understood that a machine learning model is only as good as its trained data, and building a model using only HPA axis response-based stressors will likely create a model that is not sensitive to lower forms of acute stress, which over time may influence HPA axis response. Our study aims to expand the in-lab stressors to include those that

produce a subjective feeling of stress, with no known HPA axis response, while also determining the sensors and physiologic features that are most predictive of physiologic stress.

Previous studies have demonstrated that men and women differ on how they respond to stress physiologically, behaviorally, and in self-reports [11, 52, 56] [11]. Nonetheless, the majority of prior work that build automated stress-detection models do not focus their studies on a specific target population [7, 21, 28, 43]. There is evidence that suggests that stress negatively affects pregnant mothers and the neurodevelopment of the fetus [19, 55]. Because of the intergenerational sequelae of stress, its modifiability, and the high motivation of pregnant women to provide a healthy environment for their developing fetuses [22], this high-stress population, can significantly benefit from a system that can automatically detect and eventually prevent stress. Therefore, we focus our study and testing of our framework on premenopausal and pregnant women.

Contributions:

Here, we examine an optimal micro-EMA that aligns with physiologic stress using electrocardiography (ECG), GSR, and heart rate in wearable sensors. We test a robust wearable multi-sensor suite on participants in the laboratory, followed by a real-world study using the sensor that participants are most willing to wear for long periods of time. Our specific contributions in this work are:

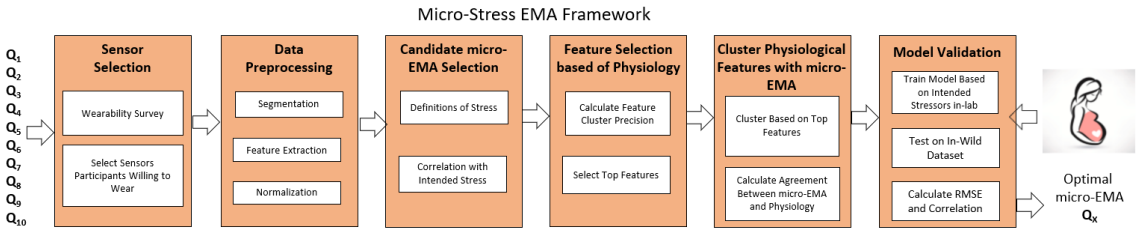

Fig. 1. Framework for determining the optimal micro-EMA question, and validation in pregnant woman wearing sensors in the field. EMA, ecological momentary assessment.

(1) We propose a methodological framework for identifying an optimal single-item micro-EMA question to use in the field to establish a measurable ground truth for stress. Such a framework can also be used for other health constructs.
(2) We define a method to identify a subset of questions that are most likely to align with the physiologic manifestation of stress.
(3) We evaluate our proposed methods in 18 young women in-lab and 17 pregnant women in the real world. We report results on a generalized machine learning model trained in-lab and in the real-world dataset. We also discuss the optimal features used in the model and provide reasoning for why the model performs poorly on some participants over others.
(4) We release our anonymized dataset of video-generated motion traces tagged with ground truth labels for use by researchers and clinicians.[1]

The framework for determining the optimal micro-EMA is laid out in Figure 1. The first step of the framework is data collection, which involves collecting both physiological and self-report data. We then select the optimal sensor suite based on participants' willingness to wear. Subsequently, we process the physiological and self-report data to extract normalized features. Using the self-report data, we reduce the size of candidate micro-EMAs based on their correlation with induced stress. Next, we apply our unbiased feature selection using Agglomerative

[1]The dataset and code is available at https://osf.io/4zajm.

clustering to select the top features that separate in-lab activities. Finally, the selected features are combined and clustered, and the candidate micro-EMAs that align best with the physiology are chosen. To further validate our framework, we build a machine learning model from in-lab data and apply that to in-wild data to show that our model built solely on physiological data can be used to select our top performing micro-EMAs.

Our work benefits patients, clinicians, and researchers studying populations in the real world with increased adherence to self-reports using micro-EMA. It also aids computer scientists in providing a framework for assessing optimal EMA questions that align with one's physiologic response to stress. This will enable stronger cross-talk between computer scientists and health professionals who care for pregnant women and their infants by identifying the optimal set of effective questions to be used in the real world when the goal is to build automated real-time detection of health constructs that align with physiology.

2 BACKGROUND AND RELATED WORK

In this section we discuss how wearable sensors and learned models have been used to build current stress models. We frame this in context of the prior work on micro-EMAs that led to the design of our framework.

2.1 Stress Models from Physiologic Signals

Hovsepian et al. [28] proposed cStress, a model that uses a wearable sensor suite to capture respiratory inductive plethysmography (RIP) and ECG from participants performing in-lab stressors to build a model that detects stress. The in-lab study involved a "socioevaluative" challenge (preparing and delivering a speech) and a cognitive challenge, which were both portions of TSST. Their results when predicting stress showed a high recall rate (89%) and low false positive rate (5%), albeit the only other non-stress-inducing activity conducted in-lab was resting. They hypothesized the high recall they obtained was mostly due to lack of diversity in the stress levels of the activities performed during the study. One factor that differentiated cStress from other stress studies was the authors test of the system in free-living populations (20 participants for 1 week). Hovsepian et al. reported a median F1-Score of 71% across the 20 participants and a median F1-Score of 75% across 26 participants. Hovsepian et al. used a variation of the perceived stress scale for ambulatory settings proposed by Cohen et al. [11] and used by Plarre et al. [43]. The scale included five questions that asked the subject how "cheerful," "happy," "angry/frustrated," "nervous/stressed," and "sad" the subject was; responses ranged from 1 to 6. Plarre et al. [43] proposed a framework for determining stress similar to Hovsepian et al. and used a similar in-lab setting with structured activities, capturing the same data using RIP- and ECG-based wearable sensors to build a physiologic stress model. The authors argued that the subjectivity of self-report will always limit its use in the real world. They considered perceived stress as a hidden state in a hidden Markov model (HMM), where the physiologic model generates observables used to infer the hidden state in the HMM. The main difference between the works of Plarre et al. and Hovsepian et al. is that the former did not directly use self-report as ground truth but instead built three separate models: a physiologic stress model, a self-reported stress model, and a perceived stress model. Plarre et al. [43] uses induced stress as ground truth for building these models. That is, if the in-lab activity was meant to cause stress, they labeled it as such. By building these three models, Plarre et al. put less weight on self-report to determine stress because of its subjectivity. We extend on both Plarre et al. and Hovsepian et al. to identify a micro-EMA that most aligns with one's physiologic stress.

Sano et al. [47] focussed on finding the physiological features and signals that associate well with perceived stress. The authors combined the use of an accelerometer and skin conductivity sensor in conjunction with some variables like phone usage, text messages, and stress surveys throughout the day. They then used these features to classify whether the participant belonged to the high stress or low stress group, which was determined using perceived stress scale (PSS) scores. One interesting observation was their method for determining important features, which was performed by finding the features that correlated with perceived stress.

2.2 Stress Models from Other Signals

Two common signals used in stress detection are audio and video. Both Chang et al. [6] and Lu et al. [36] proposed models using audio signals from smartphones. StressSense uses a smartphone to record a subject in a laboratory environment similar to TSST but also has the subjects perform unstructured tasks in real-world situations. The results of the model vary depending on what was used as the test and train set. In the study by Chang et al. [6], when training on in-lab and testing in-the-wild, the F1-Score ranged between 38.9% and 47.7%, respectively, but when training on in-the-wild and testing in-lab, the F1-Score ranged between 62.2% and 77.5%. Chang et al. [6] proposed Ammon, a framework for processing audio signals and then applying affect and mental health recognition.

Most researchers who use video signals for stress detection are actually using facial recognition to determine stress. Gao et al. [24] used facial recognition classification models built from two facial expression databases. Then a model was built using a support vector machine (SVM)-based one-vs-all voting mechanism where the classes were the different emotions. The authors found that stress was a combination of anger and disgust. Another similar study, but for a different application, was proposed by Dinges et al. [17] who used facial expressions to determine stress during space shuttle flights. This method of detecting stress is challenging to implement in a truly free-living environment but can work in a car or cockpit, as it is possible to have a camera always facing the subject while seated and in a confined space.

The main concern with using both audio and video signals in stress detection are the privacy concerns of subjects and bystanders [15]. Many individuals find it concerning to know that their voice and the voices of people around them are being recorded [30]. Some of these privacy concerns, specifically recording those around the participant, can be circumvented by using an egocentric wearable camera to record video [27]. However, being video-taped can affect a subject's naturally occurring behavior [1]. For instance, they might not want to say something that could be used out of context or put them in a negative light.

2.3 Ecological Momentary Assessment

One reason Plarre et al. [43] argued that using self-report to determine ground truth of stress is not realistic is that it can only establish truth for a short amount of time. Additionally, having a subject answer multiple questions multiple times a day is highly burdensome and is prone to low participant response. Plarre et al. attempted to mitigate this problem by building a model using in-lab stress induction and labeling stress episodes independent of the subject's self-report. The authors then built separate models based on both the self-report and physiologic data that was able to predict perceived stress. While this was able to accommodate the variability in the subjective reporting of stress, it did not solve the burden associated with self-reporting stress in a real-world environment. Ponnada et al. [44] performed a study on the association of burden and self-reports with EMAs and found that micro-EMAs can significantly alleviate some of the problems participants have with low adherence. Traditional EMAs use a smartphone [26] to prompt a subject to complete a self-report measure; usually the subject is prompted 5-10 times a day. Conversely, a micro-EMA is a single question that is asked on a smartwatch or a smartphone, in which the subject can quickly and easily respond. Ponnada et al. showed that one can prompt a subject with a micro-EMA up to 8 times more often than a traditional EMA, reducing the burden felt by participants and, as a result, increasing adherence to self-report. One disadvantage of the micro-EMA is the inability to assess multiple feelings, which creates a challenge to addressing complex constructs such as stress. However, we have designed a framework that will enable researchers to identify the optimal single micro-EMA that can be used to properly predict physiologic stress by selecting a self-report that best aligns with ones physiology.

2.4 Stress in Pregnant Women

Previous studies have demonstrated that men and women differ on how they respond to stress physiologically [52], behaviorally [56], and in self-reports [11]. Kelly et al. [32] studied the differences in physiologic response to the TSST. The authors found differences when reporting negative affect, specifically with women reporting higher negative emotion compared with men, without a major difference in physiology (cortisol, heart rate). In the American Psychological Association's annual report on stress [2], there were many differences between the genders with regard to stress, including how each gender prevents, reduces, treats, and reports stress. It is also evident that the two genders react differently to stress. In fact, a study was conducted that showed that women are more likely to stress eat than men [51].

There is evidence to suggest that stress during pregnancy can affect pregnant mothers, as well as fetal neurodevelopment. Dole et al. [19] conducted a study on more than 2000 pregnant women and found that high levels of negative and stressful life events were associated with preterm births (< 37 weeks). Other studies have concluded that there is a relationship between prenatal stress and fetal behavior [55], and there is also evidence to support the relationship between prenatal stress during pregnancy and the long-term effects on the child. Our study focuses on premenopausal women in early adulthood (18-34 years) as a proof of concept for our framework. With the varying stress responses of men and women and then women and pregnant women, we believe that prior to testing a stress model in the field, determining the optimal micro-EMA should be identified according to the target population of interest. But due to concerns of inducing stress in pregnant women in-lab, we first tested the stress induction on premenopausal women, and then perform the in-the-wild test on pregnant women. Pregnant women have the potential to significantly benefit from a system that detects normative variation in stress during pregnancy, and in the future a timely stress-reduction intervention can be applied. Pregnant women are also known to be a high-stress population worthy of dedicated study [48].

3 STUDY

In order to collect micro-EMA responses in varying stress levels, we developed a controlled study in an in-lab environment in which participants performed stress-inducing and non-stress-inducing activities. Recruitment for the study was accomplished through posted flyers on a college campus in the Midwest, as well as at a nearby university hospital. Participants could be included if they were female (not pregnant) and between the ages 18 and 35 years.

3.1 Devices

During the study participants wore a suite of three sensors (Figure 2). The sensors included (1) the BiostampRC,[2] a flexible wearable patch for gathering raw electrocardiograph (ECG) data at 250 Hz; (2) the Polar H7,[3] a heart-rate monitor worn around the chest through a chest strap providing heart rate estimates at a 1-Hz sampling rate; and (3) the Neulog GSR module,[4] which uses two probes connected to the participant's fingers to capture GSR at 2 Hz. GSR is a measure of skin conductivity, which varies based on how much the subject perspires at the body location (i.e., index and middle finger).

3.2 Structure of the Study

Our study was designed to better understand the relationship between an individual's physiology and self-perceived stress. Multiple studies [28, 36, 43] have attempted to build a machine learning model that detects perceived stress arousal and then apply their in-lab findings to real-world data. We argue that this approach

[2]https://www.mc10inc.com/our-products/biostamprc
[3]https://support.polar.com/us-en/support/H7_heart_rate_sensor
[4]https://neulog.com/gsr/

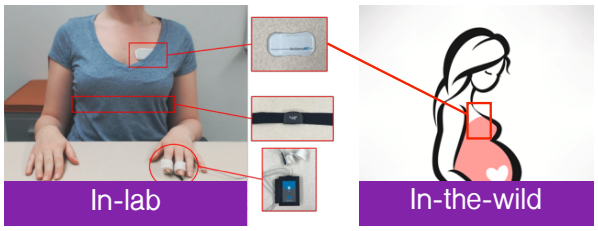

Fig. 2. Devices used in our study: Neulog GSR module, Polar chest strap, and BiostampRC. GSR, galvanic skin response.

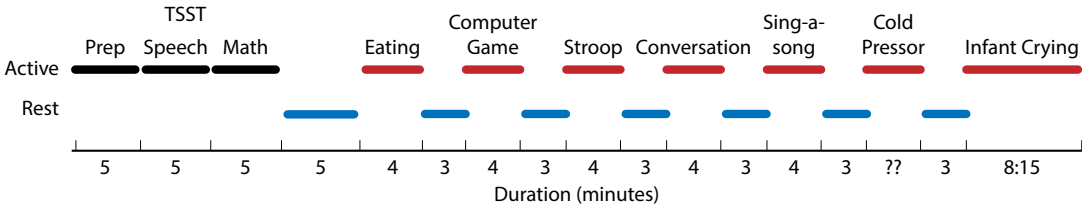

Fig. 3. Example distribution of activities during testing. Black lines denote sections of TSST; red lines are non-TSST activities; blue lines are rest periods.

can only capture extremely stressful situations, while ignoring moments of acute stress. Based on initial results, we found that we could predict physiologic arousal of stress with 90% accuracy using an SVM model with a radial basis function (RBF) kernel on features selected using CfsSubsetEval when only using TSST and rest-based activities. However, when we applied this model to our entire dataset with varied stress inducers, the accuracy dropped to 62%. Therefore, we determined that varying the stressors in-lab was essential to constructing a robust model. Our study had participants perform certain activities and then asked them to fill out a self-report questionnaire following each activity; the self-report contained 12 questions and was filled out before the start of their next activity. The in-lab study lasted 1 hour and 44 minutes on average, including preparation time between activities and time for self-report completion.

For each participant the study began with a psychologic stressor, the TSST, a well-established stress-inducing technique [33] comprising three parts: a speech preparation phase, a speech delivery phase, and a mental arithmetic test. The participant was then asked to perform subsequent activities that induced physical and mental stressors (see Table 1) in random order to minimize carryover effects. Each activity was followed by a rest period, with the exception of the final activity. Each rest period lasted 3 minutes and was accompanied by turning off of lights and playing of a soothing video with audio ranging from natural scenery (ocean waves hitting the shore) to calming classical music. Figure 3 shows a sample distribution of activities. Duration of the cold pressor test, which induced physical stress, varied by participant, as each participant was instructed to leave his/her hand in the bucket of ice water, although they could remove it at any time. Each of the in-lab stressors were selected based on a literature review of different stress-induction methods; however, while they are generally accepted to be stressful or not, they will still vary in their effect from one participant to another. As a result, upon completion of each task, participants were asked to self-report their stress level for each of the three phases (see Table 2 for a full list of self-report questions). At the end of the in-lab study, we asked participants to report their willingness to wear each device to inform our real-world study on pregnant women.

Table 1. Description of each activity completed during our in-lab test. Red activities are validated/intended stressors, while black activities are unvalidated/unintended stressors.

Activity	Description
TSST [33]	5 min: Preparing for speech 5 min: Giving speech about self 5 min: Performing mental subtraction
Eating	4 min: Eating from an assorted set of foods
Computer Game [3]	4 min: Playing online car racing game
Stroop [12]	4 min: Typing the first letter of the color that appears on a laptop screen
Conversation	4 min: Unstructured conversation with the researcher
Sing-A-Song [4]	4 min: Singing any song for the full duration
Cold Pressor [49]	≈ 1 min: Inserting hand in ice bucket for as long as participant can.
Infant Crying [39]	8:15 min: Listening to a 45-sec recording of an infant crying with two 3-min rest periods between the recordings
Rest	3 min: Watching a peaceful video with calming music while lights are turned down

Table 2. Micro-EMAs derived from our questionnaire. Negative signs denote negative emotions, while positive signs denote positive emotions. Each question also contains the range for the response.

Activity	Description
Intended (-)	Based on whether the activity was intended to induce stress
LikertStress (-)	How Stressed were you? (0-6)
BinaryStress (-)	Were you Stressed? (yes/no)
PSS-Control (-)	Did you feel you could not control important things? (0-4)
PSS-Overcome (-)	Did you feel difficulties piling up so you cannot overcome them? (0-4)
WorriedStress (-)	How Worried were you? (0-100)
SadStress (-)	How Sad were you? (0-100)
IrritableStress (-)	How Irritable/Angry were you? (0-100)
PSS-Confident (+)	Did you feel confident in your ability to handle problems? (0-4)
PSS-Your Way (+)	Did you feel things are going your way? (0-4)
ContentStress (+)	How Content were you? (0-100)
HappyStress (+)	How Happy were you? (0-100)
ExcitedStress (+)	How Excited were you? (0-100)

3.3 Self-Reporting Measures

There are multiple methods for determining perceived stress through self-report data. Twelve questions, commonly used to assess perceived stress, were fielded to the participants both in-lab, after each activity, and in the real world. These questions are listed in Table 2. The intended stress definition is represented by a 0 for the non-stressful activities described in Table 1 (eating, conversation, and rest) and by a 1 for the stressful activities (proven factors intended to be stressful). Plus and minus signs next to the stress definitions in Table 2 represent whether that question relates to a positive or negative emotion, respectively. Some of the questions are from the multi-item PSS 4-item questionnaire (PSS-4) [11], commonly used to assess stress (questions have prefix "PSS-" in Table 2) while the remainder come from literature.

Table 3. Information pertaining the in-wild data set. For participant 18 we see the time the device was worn, but the data was corrupted during the upload. Those participants who did not upload data on day 2 were due to trouble of starting/stopping the sensor. On average participants wore the device for 12 hours and 38 minutes.

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Day 1	8:10	10:59	10:34	10:45	12:36	8:23	10:25	3:43	7:29	10:41	8:08	6:18	10:43	6:24	7:36	2:12	6:52	4:55
Day 2	3:56	6:57	3:47	0:00	3:10	4:19	5:15	5:40	5:07	0:00	4:10	5:25	10:30	5:37	0:00	8:11	0:00	8:34
Total	12:06	17:56	14:21	10:45	15:46	12:42	15:40	9:23	12:36	10:41	12:18	11:43	21:13	12:01	7:36	10:23	6:52	13:29
EMAs	6	6	5	9	4	6	6	5	6	6	6	3	6	5	5	5	6	5
PSS-Q4	2.3	3.8	1.8	4.2	1.8	3	0.7	4.8	9.8	0.8	4.8	0	1.3	3.4	2.6	0.6	3.2	0.4
Clean	82%	72%	59%	40%	34%	45%	59%	55%	49%	51%	43%	53%	44%	57%	56%	47%	28%	0%

For each method, we also established a threshold that distinguished between stressful and non-stressful activities. Many of the questions used "personalized threshold," which was defined by calculating the mean of the subject's response for a specific question across all tasks. For the negative emotions, if the participant response rating was greater than the personalized average response, then the activity was labeled as *stressful*. The opposite was true for the positive emotions.

3.4 Micro-EMA Tests in Pregnant Women in the Real World

To test the in-lab model on real-world data, we recruited 18 pregnant women (in their first or second trimester for stability). Stress induction was not performed on the pregnant women; rather we analyzed the model developed in-lab in pregnant women in a real-world setting. Based on in-lab participant feedback (see Section 5.1), we determined that women would be most willing to wear the BiostampRC flexible sensor in the real world compared with the Polar H7 chest strap or the NeuLog sensor. Each pregnant woman visited the laboratory, where they were taught how to use the device and how to respond to stress-based EMA questions throughout the day on their smartphone (which required response to $\geq$5 prompts). The EMA prompts were sent via a text message. Each prompt contained several questions related to their stress and emotional state in the past hour, covering each micro-EMA outlined previously (Table 2). They then wore the device in the field for the remainder of the day. On the day of the second visit (the following day, unless the participant had a scheduling conflict), the participant wore the device in the morning and started recording until they returned to the laboratory, where they completed a wearability survey. Participants wore the device for roughly 12 hours in the real world. This study was approved by an institutional review board, and all participants provided written informed consent. Few participants had connectivity problems during the end-of-day data upload, which resulted in some data loss. Details about the amount of data collected can be found in Table 3.

4 METHODS

Existing literature focuses on inducing stress in laboratory settings to build a stress model that works in the real world. But no study identifies a method for determining the optimal set of self-report questions that are most correlated with the physiological stress response. We propose a framework for determining an optimal micro-EMA that aligns with physiologic stress in a given population using both in-lab and real-world data. To do this, we analyzed how well each EMA question maps onto physiological manifestations of stress represented by features collected from the wearable sensors. In doing so, we determined the set of EMA questions that were most likely to align with physiology. We began by clustering significant features extracted from the wearable signals (representing the physiology of the participant) and then calculated the agreement between the resulting

clusters and each EMA question/definition. The various stages of this framework are defined in Figure 1. We began by selecting candidate micro-EMA questions using point-biserial correlation. Next, we ran an unsupervised feature-selection algorithm using clustering 10 times and selected any feature that was found to separate the data set by stress/non-stress activities.

We then perform feature selection using a two step approach. First, we cluster the data and see which features provided the greatest cluster separability. We tested several k values for clustering and selected the value of k which shows the greatest improvement in cluster separability. Next, we determined how well each micro-EMA aligned with each cluster and selected the micro-EMAs that aligned best by calculating cluster separability and Cohen's Kappa and selected the top three micro-EMAs. We then applied correlation-based feature subset selection (CfsSubsetEval) on the features to narrow down the features and build a machine learning model using the optimal features from the data in-lab. We compared against several machine learning classifiers previously shown to be effective at predicting stress. We further validated the model in the field, and assessed how well the top candidate micro-EMAs aligned with our model (trained and built from induced stress in-lab) on in-the-wild data. Finally, we selected the micro-EMA that had the highest accuracy with the in-field data.

4.1 Sensor Selection

Following the in-lab and real-world studies, we asked participants to fill out a sensor wearability survey. The survey included multiple questions on comfort of each device, style, and general concerns with the devices. We also asked about willingness to wear the devices for extended periods of time. During the real-world test, participants wore only the BiostampRC, but they were asked to wear both the Polar H7 chest strap and the Empatica E4 while in laboratory before completing the surveys.

Using the survey and the sensors applicability to predict stress, we selected the optimal subset of sensors. We first looked at three factors: (1) if the participants were willing to wear the sensor for 30 days, (2) if the device was portable, and (3) if the device was comfortable. Based on the results, we tested whether the device or devices could be used to reasonably measure stress.

4.2 Data Collection and Preprocessing

To train a potential stress model, we used sensor data collected from in-lab data from 18 participants. Six participants had partially incomplete data due to technical problems. Two had an error when uploading the ECG data using the provided tablet. Two had incomplete heart rate data caused by Bluetooth connection loss between the device and the smartphone, and two had incomplete self-report and ECG data. Our class labels were stress (1) and no-stress (0) based on the proven intended stressors in-lab.

4.2.1 Segmentation. Before extracting features from the data, we segmented the data using fixed-time subdivisions of 1 minute, with a sliding window of 50% overlap in segments to detect stress at a minute level. The class label associated with each segment was obtained using the intended stressors. For instance, since the speech activity from TSST is a proven intended stressor, then each minute corresponding to the speech is labeled as stressful.

4.2.2 ECG Processing. ECG data was cleaned using a noise removal algorithm [58] tested on the BiostampRC both in structured activities performed in-lab and unstructured activities performed in the real world. To remove noise in the ECG data due to motion (high-intensity activity) and adhesive artifacts, a band-pass filter removed all data >200 Hz and <0.6 Hz. Data were then segmented into fixed-length windows, and an ensemble model combining SVM and a feed-forward neural network classified each window as clean or noisy. Figure 4 shows an example of an ECG signal that contains both clean and noisy data. The first and last 10 seconds are noisy due to connectivity issues. The middle portion highlighted in green represents clean data which can be salvaged and

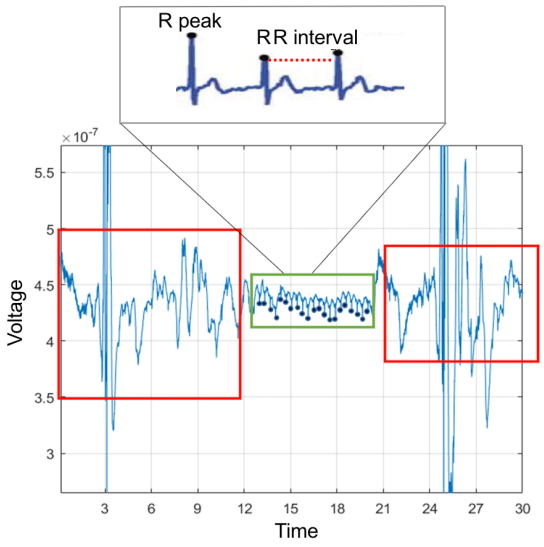

Fig. 4. BiostampRC signal with noise (red box), and clean data (green box).

Table 4. Features Extracted. Once IBIs are found in the clean data, we extract both heart rate variability based features and statistical features (non-heart-rate variability). RMSSD, root mean square of successive RR interval differences; SDSD, standard deviation of successive RR interval differences (short-term variability); pNNX, percentage of successive RR intervals that differ by more than X ms; IQR, interquatile range.

Signal	Features
Heart rate variability	Low freq. energy (0.1-0.2 Hz) Medium freq. energy (0.2-0.3 Hz) High freq. energy (0.3-0.4 Hz) Low:high freq. energy ratio Variance, quartile deviation RMSSD, SDSD pNN20, pNN50
Non-heart-rate variability	Mean, median, mode Minimum, maximum, range Root mean square (RMS), zero cross Kurtosis, skew, IQR 20^{th}, 40^{th}, 60^{th}, 80^{th} Percentile Count>mean, count<mean

used for stress detection. The dark blue dots indicate clean R-peaks, and the top graphic shows the QRS complex alongside RR IBI. After the signal was cleaned, IBIs were identified using the Pan-Tompkins algorithm [40].

We compared the performance of features extracted from BiostampRC only (IBI features based on ECG), Polar H7 chest strap heart rate, and NeuLog (GSR), as well as the features from the combined sensor streams (all sensors). We determined how well the IBI features performed in relation to the other sensing modalities. However, we highlight that if participants are not willing to wear the sensor in the real world, then that should take precedence over the information gained from including the sensor in the model.

4.2.3 Feature Extraction. The features extracted from the device suite are presented in Table 4. Using the three devices, we extracted multiple features known to be useful for predicting stress, combining statistical and frequency-based features.

4.2.4 Normalization. The final step of data preprocessing was normalizing the data. A unity-based normalization approach was selected to keep all values of a feature in range [0, 1], as feature scaling is a necessary step prior to applying machine learning algorithms. However, due to the unique physiologic manifestation of stress in each individual (heart rate intervals may vary among participants), we normalized each individual's data separately.

4.3 Candidate micro-EMA Selection

The first step of the micro-EMA framework was to select candidate questions for our micro-EMA. A micro-EMA only allows for a limited number of questions to be asked (ideally 1 question), and thus we extensively reviewed the questions that have been studied in the literature. To identify a subset of the candidate questions, we calculated

the point-biserial correlation between each stress definition (see Table 2) and intended stress (i.e., stressors that have been verified in literature including TSST, sing-a-song, cold pressor, and Stroop). We then selected the candidate micro-EMA questions that yielded moderate to strong correlations with intended stress. We also analyzed a few micro-EMA questions that resulted in the greatest negative correlation with stress to determine how well positive and negative emotions aligned with physiologic stress.

4.4 Feature Selection Based on Physiology

The main objective from this step in our micro-stress EMA framework was to identify physiologic features that can separate stress and non-stress situations in in-lab study. Thus, to achieve this goal, each feature from the selected sensors dataset was used by itself to cluster stressful and non-stressful activities (determined by the intended stressors) with k-means clustering. Those features that showed good separability as in the examples of mean_IBI and min_IBI in Figure 5 were selected for the feature set. Features like max_IBI, in the same figure, which show poor separability between stressful and non-stressful instances were eliminated as candidate features in our final model.

In order to quantify good/poor cluster separability, we calculated cluster separability, a metric adapted from Huynh et al. [29]. The metric takes into account the purity of each label (e.g., stress/no stress) within the clusters and calculates a total cluster separability across labels. Labels with high cluster separability are well separated from other labels. To calculate cluster separability, we first calculated the ratio of each label in each cluster:

$$p_{i,j} = \frac{|C_{i,j}|}{\sum_j |C_{i,j}|}, \qquad Eq.1$$

where $p_{i,j}$ denotes the separability for the jth label in the ith cluster, while $C_{i,j}$ denotes the total number of the jth label in the ith cluster. We then calculated the separability of each class label P_j across all clusters i:

$$P_j = \frac{\sum_i p_{i,j}|C_{i,j}|}{\sum_i |C_{i,j}|}, \qquad Eq.2$$

Cluster separability is also used to determine the optimal k number of clusters when determining the performance of each of our candidate micro-EMAs in Section 4.5.

4.5 Cluster Physiologic Features with micro-EMA

In this step of the micro-stress EMA framework, we used features selected from the previous step to determine candidate perceived stress EMAs that align best with physiology. We begin by applying the Agglomerative clustering algorithm [57] based on Ward's minimum variance method [57]. The number of clusters tested ranged between 2 and 16, representing the different activities. We then analyzed the relationship between each micro-EMA question and the physiologic data. If the activities were clustered physiologically into two clusters, one representing stress and the other non-stress, then we would have identified a micro-EMA that best aligned with the two classes, where all the activities labeled as stressful were in one cluster and the non-stressful activities in another.

We clustered using the intended stressors and non-stressors from in-lab and determined how the clusters align with the different micro-EMAs. Two of the evaluation metrics informed which of the micro-EMAs best aligned with physiology. These metrics included cluster separability, described in Section 4.4, and Cohen's kappa [10]. Cohen's kappa measures inter-rater agreement and takes random agreement into account when calculating agreement. The first rater is defined by associating each cluster with each micro-EMA. For instance, if there were more stressful episodes in one cluster, we identified it as the stressful cluster and determined its agreement with the micro-EMA samples assigned to that cluster. If there were more non-stressful samples in one cluster, we labeled it as non-stress and determined its agreement with the rest of the samples in that cluster. We then

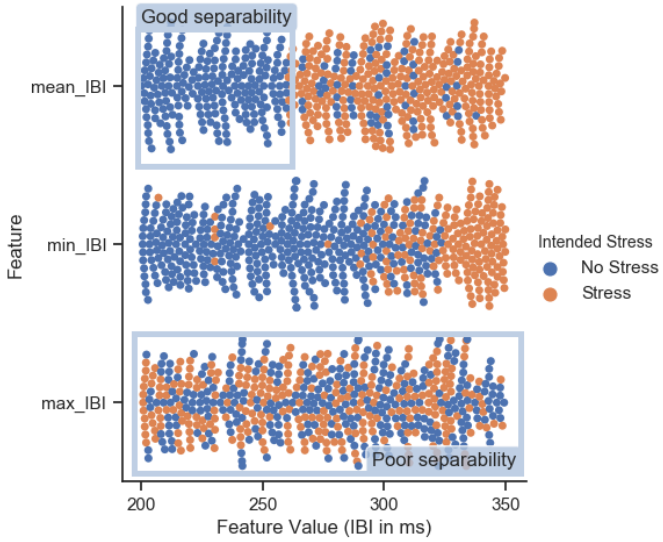

Fig. 5. Example of cluster separation with 3 different features. Max_IBI feature shows poor separability, while mean_IBI and min_IBI features show more precise clusters with good separability.

calculated cluster separability to identify the separation across the stress and non-stress samples as defined by each micro-EMA (instead of the intended stressors) to see which best aligned with physiology, thereby finding the micro-EMA that can be used to label physiological data in-wild.

4.6 Model Validation

4.6.1 Training Model from In-Lab Data. Looking at literature we selected five machine learning methods that are widely used for a similar application.

- **Support Vector Machine (SVM):** A supervised learning machine used in two-group classification [13]. An SVM will produce a hyper-plane based on the data provided that best separates the classes. We can tune several hyper-parameters to optimize the SVM. One is the kernel function, which is used to manage the high-dimensional data. For our SVM we selected the radial basis function (RBF) kernel, which also required us to define a γ value, which defines how far the influence of a single training sample reaches. Another parameter is the soft-margin parameter C which essentially determines how rigid the hyper-plane must be, a large value of C will allow for very little misclassification when outputting the hyper-plane. To determine the best value for C and γ we applied an exhaustive grid search choosing the best combination of C and γ that optimize our model in terms of F1-Score [28, 43, 53].
- **Decision Tree:** A tree-like structure, at each layer there is a condition, and the branch taken is determined by the data. This continues depending on the layer of the tree, until it reaches a leaf node that determines where it is a stressful or non-stressful event [38, 43, 53]. We set the max decision tree depth as 10.
- **AdaBoost:** A machine learning algorithm [23], and in our case it works in conjunction with a decision tree. At every layer of the decision tree, AdaBoost proposes a weak classifier that might not be accurate. It then applies weights to each of the data points depending on whether they were classified correctly or not. Then at each layer the algorithm places emphasis on classifying certain data points correctly depending on

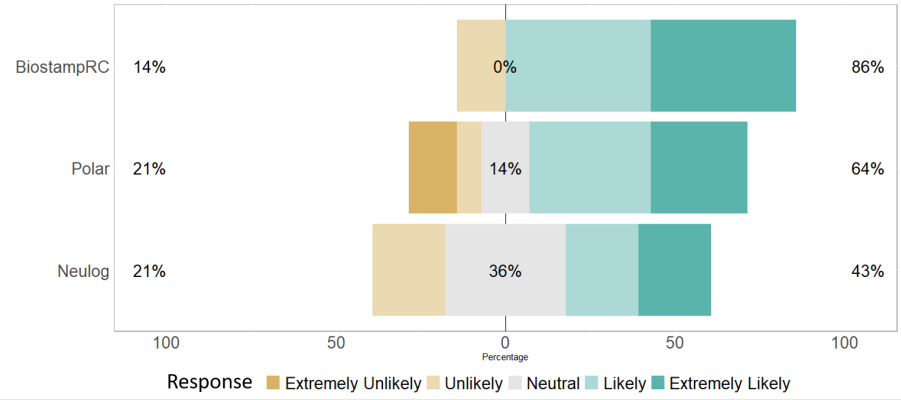

Fig. 6. Based on the question "If paid $100, would you be willing to wear the sensor for 30 days?" 1-extremely unlikely, and 5-extremely likely. The percent on the left indicates the percentage of participants who responded as extremely unlikely or unlikely to wear the device, while the percent in the middle denotes the percentage of participants who responded neutrally, and the percent on the right is the percentage of participants who responded as extremely likely or likely.

their weights. This continues until all of the data points are correctly classified or it reaches a threshold for the number of layers (defined by the user) [43]. We use decision stump as the base classifier.

- **Naive Bayes:** A classifier based on conditional probability. For each feature we calculate the probability that a condition is met given that the sample is classified as stressful or not. Based on these probabilities given to a data point, we calculate the probability that it belongs to either of the classes [53].
- **Neural Network:** Neural Network is a machine learning algorithm that employs an interconnected group of nodes (or neurons) to model the non-linear relationship between input features and the label. We use a 3-layer neural network with 10 neurons and select Sigmoid as the activation function.

4.6.2 Test Model on In-The-Wild Data. The model was built from in-lab data using structured activities, with intended and non-intended stressors as ground truth. We then tested the model in a real-world setting. Since we only captured ECG from the BiostampRC in pregnant women, our model was built using the ECG-only (i.e., IBI features only) dataset. The selected micro-EMAs are used as the class labels for the test set. We report the accuracy during stressful and non-stressful periods. We also assessed varying window lengths prior to the stress self-report between 1 and 60 minutes in order to assess how proximity to the self-report effects accuracy.

5 RESULTS

5.1 Sensor Selection

A total of 14 participants (63.6%) responded to the wearability survey. When asked if a participant would be willing to wear the device for 30 days if they would receive $100 for doing so (1=extremely unlikely, 5=extremely likely), 86% of participants said they would be "likely" or "extremely likely" to wear the BiostampRC compared with 43% of participants for the NeuLog and 64% for the Polar H7 chest strap. Responses were similar during the real-world portion. Across all 18 participants who participated in the real-world study, when asked "To what degree would you be likely to wear the device during you daily activities for 12 weeks?" the mean response for both Empatica E4 and BiostampRC was 3.44 and Polar H7 chest strap was 3.06. Although, participants score the Empatica E4 and BiostampRC with similar scores, a significant source of apprehension with wearing the BiostampRC was not knowing how it would affect them or their unborn child. For example, one subject who

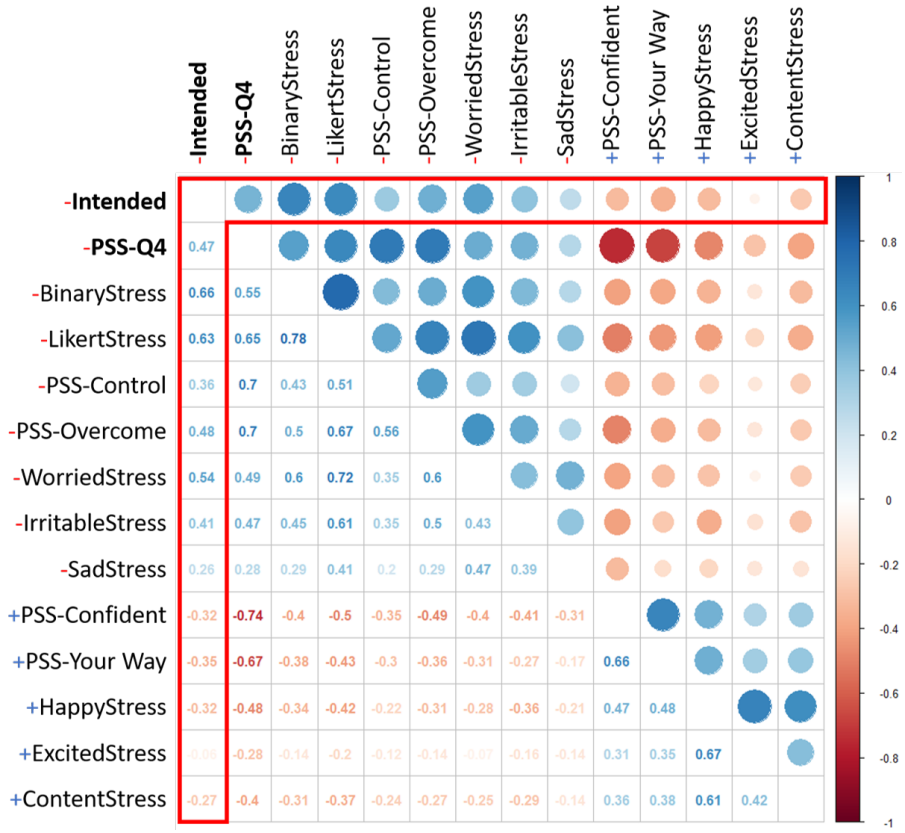

Fig. 7. Inter correlation between all of the self-report questions and Intended stress. The correlation between intended stress and the self-report questions is outlined in red.

reported that they were extremely unlikely to wear the BiostampRC, reported she wanted to be sure it would not affect her baby. When asked how comfortable the device was (1=extremely uncomfortable, 5=extremely comfortable), the average response was a 4.28 for the BiostampRC and 3.33 for the Empatica E4 and Polar H7 chest strap. Based on these results we chose BiostampRC as the sensor we would test in the wild.

5.2 Micro-EMA Candidate Selection

To identify candidate micro-EMAs, we measured each micro-EMAs correlation with intended stress and determined four candidates that exhibited moderately positive correlation: BinaryStress (r = .66), LikertStress (r = .63), PSS-Overcome (r = .48), and WorriedStress (r = .54). Despite having weaker correlation, we identified two negatively correlated micro-EMAs to compare their potential for stress predictability to the positively correlated micro-EMAs, PSS-YourWay (r = −.33) and HappyStress (r = −.32), to determine the ability of positive emotions to predict stress. Point-biserial correlations for all 12 micro-EMAs are shown in Figure 7. Sano et al. [47] proposed a similar method for determining important features for predicting perceived stress. Instead we treated induced stress as our ground truth for physiologic stress and determined the questions that most correlated with physiologic stress.

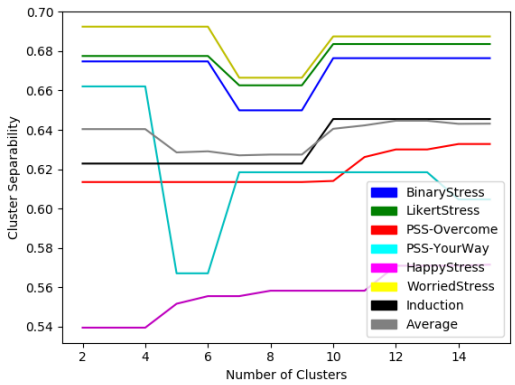

Fig. 8. Cluster separability of each selected micro-EMA and induced stressors (Induction) where k is between 2 and 15.

Table 5. Evaluation metrics for determining agreement between clustered physiology and self-report using ECG only.

Stress Definition	**Cohen's Kappa**	**Cluster Separability**
BinaryStress	0.32	0.67
LikertStress	0.34	0.68
PSS-YourWay	0.21	0.61
PSS-Overcome	0.00	0.66
WorriedStress	0.34	0.69
HappyStress	0.07	0.54

5.3 Feature Selection by Clustering

Using the features extracted from the BiostampRC ECG signal, we calculated the cluster separability for each feature using intended stress as our class label. This returned two values of cluster separability P_j in terms of stress and non-stress. We averaged the two and selected the 10 features that returned the highest averages. The selected features were mean, min, range, mode, low frequency energy (LF), 40^{th} percentile, 60^{th} percentile, 80^{th} percentile, standard deviation of successive RR interval differences (SDSD) representing short-term variability, and root mean square of successive RR interval differences (RMSSD).

5.4 Micro-EMA Selection

Figure 8 shows the cluster separability against the different definitions of stress as a function of the number of clusters, k, which is used to determine the optimal number of clusters. We selected the optimal k by calculating the average cluster separability across all of the selected micro-EMAs and observed that the greatest increase in cluster separability occured as k went from 3 to 4 clusters. We did not look at the best performing k because we assumed that as we increased k, the results would improve steadily due to over-fitting. Moreover, fewer clusters would be more meaningful for interpretation. Table 5 shows the cluster separability and Cohen's kappa across the six candidate definitions. When assessing significance of Cohen's Kappa 0–.20 is considered slight, .20–.40 is fair, .40–.60 is moderate, .60–.80 is substantial, and .80–1.0 is almost perfect agreement [35]. We also expect less agreement due to the subjectivity of stress. The two definitions with the highest cluster separability and reasonable Cohen's kappa were WorriedStress (kappa = 0.34 and cluster sep. = 0.69) and LikertStress (kappa = 0.34 and cluster sep. = 0.68); each outperformed PSS-Q4, the multi-item scale. We determined that LikertStress outperformed WorriedStress in cluster separability of the stress class label.

5.5 Validating Model

Before building our models we selected features using correlation-based feature selection (CFS) [25]. CFS works by selecting features that correlate well with the class label, but do not correlate between features. When applying this to the LOSO model, we selected features for each training set used, and when testing on the in field data, we used the entire in-lab data to select the optimal features.

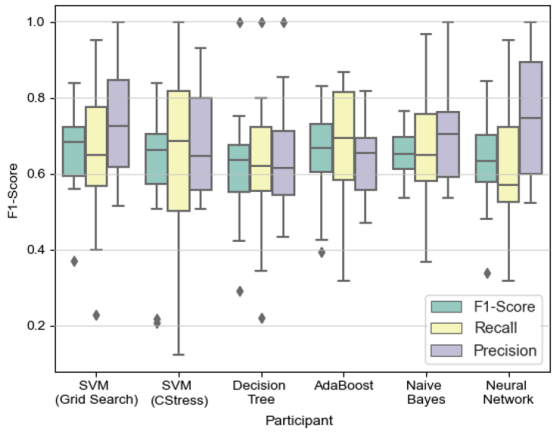

Fig. 9. **In-Lab:** Box plot is based on each participant's precision, recall, and F1-Score. We use a leave-one-subject-out (LOSO) cross validation approach. SVM (Grid Search) C=107 and γ=0.001; SVM (CStress) C=724.077 and γ=0.022097.

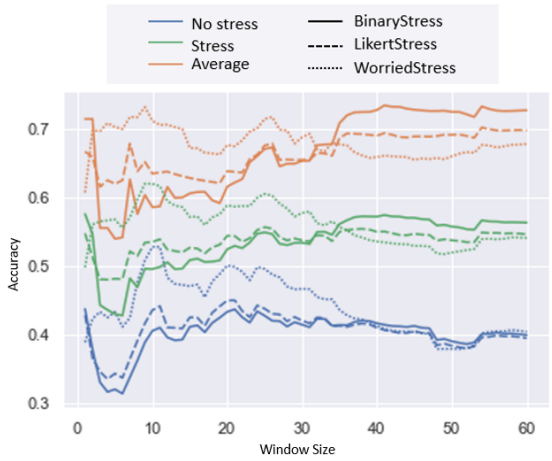

Fig. 10. **In-Wild:** Accuracies for predicting stress and no stress when varying the window size of our class label. We also show the average accuracy between stress and no stress events. Each of the accuracies are split by our three best performing definitions of stress denoted by the type of line.

Figure 9 shows the results of using the different models. We found that using SVM with the exhaustive grid search performed the best, with an F1-Score of 70%, while all of the other models performed similarly. Next we applied our optimal model (i.e., the SVM-Grid Search method in Figure 9) on in-field data and calculated the accuracy of stressful and non-stressful activities. We did this using the top 3 performing candidate micro-EMAs as our ground truth in the wild. We also varied the window lengths of the ground truth. For instance, the question was asked in terms of the past hour, but due to recall bias the answer may not have manifested during the entire 1 hour prior to the self-report. This is why we assessed multiple window sizes, as even though we were asking about their emotions in the past hour, their response may have only been representative of the past 10 minutes. Figure 10 shows the accuracy of varying the window lengths while using the three selected micro-EMA questions. We show that when using short length windows, WorriedStress returns the highest accuracy (which seems to be the closest to the self-report), but when using longer length windows, LikertStress and BinaryStress perform better.

To determine whether certain participants were adversely affecting our trained model, we examined the positive and negative F1-Scores shown in Figure 11, which show the results when applying a leave-one-subject-out approach to our in-lab dataset. Based on these results we selected two of the best performing and two of the worst performing participants and looked at how the most significant selected features differed during stressful and non-stressful activities. The results are shown in Figure 12. Figure 12 shows that participants 7 and 11 had greater variability between stressful and non-stressful activities across the different features. However, participants 1 and 5 showed greater similarity between the top performing features, except from the max IBI value (for participant 5), which shows unusually low values for stress, suggesting the activities may not have been stressful to the participant. For participants 1 and 5 we show that the max IBI value was lower for stressful than non-stressful activities, which could be one of the reasons why those participants exhibit poor predictability.

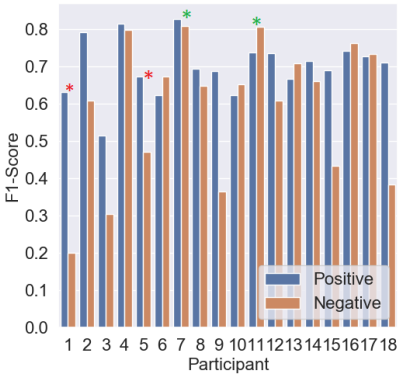

Fig. 11. **In-Lab:** F1-Scores reported for the positive (stress) class and the negative (non-stress) class when using a LOSO approach.

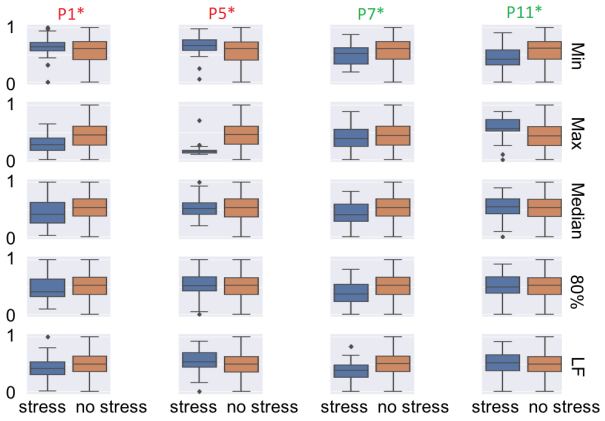

Fig. 12. **In-Lab:** Boxplots of the normalized feature values for the two worst performing participants and the two best performing participants based on Figure 11. The columns are the different participants, while the rows are the most important features when using CFS.

6 DISCUSSION

Our framework of selecting a micro-EMA for detecting stress sheds light on the challenges associated with detecting stress in real-world situations.

6.1 Micro-EMA Candidate Selection

We selected candidate micro-EMA questions by assessing correlations with intended, known in-lab stressors. Based on our results, four questions showed moderate-level positive correlation with intended stress, and two questions showed weaker negative correlation with intended stress. The questions selected were: BinaryStress, LikertStress, WorriedStress, PSS-Overcome, HappyStress (negative correlation), and PSS-YourWay (negative correlation). While researchers would typically stop at this point and select the question with the highest positive or negative correlation, this may not be a sufficient metric for selecting the micro-EMA. Our results showed that BinaryStress resulted in the strongest correlation with the intended in-lab stressors; however, Cohen's kappa suggests little agreement. WorriedStress returned the highest Cohen's kappa coefficient, but it did not perform well when calculating cluster separability for stress, suggesting it is better at determining non-stressful events but not necessarily stressful events. However, despite a high correlation in in-lab data, BinaryStress was also found to be inferior to LikertStress in predicting stressful events, suggesting that binary questions are not ideal in distinguishing between stressful and non-stressful events.

6.2 Clustering Analysis

Based on the cluster alignments, we selected LikertStress, WorriedStress, and BinaryStress as the micro-EMA questions with the best alignment with physiologic stress. Furthermore, as shown in Figure 8, we also decided to cluster based on a cluster size of four, suggesting that stressful and non-stressful activities may be divided into more than two clusters (with some stressful activities varying in intensity and type).

6.3 Validation

The results for testing the different models found in literature are shown in Figure 9. The best performing model is the SVM which is optimized using grid search. We found the optimal hyper-parameters to be a C value of 54 and a γ value of 0.0002. We confirmed our finding by using our optimal model on our real-world dataset. We selected WorriedStress as the micro-EMA that showed the most promise, as it performed well during short window lengths in close proximity to the self-report (see Figure 10). WorriedStress had the highest accuracy (≈0.7; stress) when looking at shorter length window sizes, while BinaryStress had a higher accuracy for the longer window size. LikertStress performed most consistently across all of the window sizes.

In Figure 12 participants 1 and 5 are those that did not perform well based on Figure 11, and for the majority of the features, there is little difference between stressful and non-stressful activities outside of max for participant 5. This suggests the reaction between the intended stressors did not manifest physiologically. This may be because certain activities did not affect the participant in the intended manner, and the participant may be better suited to handle stress or does not manifest stress physiologically in a detectable manner given our sensor. When analyzing the participants that performed well, we see clear differences between stressful and non-stressful activities for more than one feature. Participant 4 showed a clear difference between the activities for features min, median, 80th percentile, and low frequency energy, which may be why we see a high F1-Score when predicting this participant's stressful minutes.

Hovsepian et al. [28] reported that the memory of stress may persist in the mind of the participant, and so there may be a variable lag between the occurrence of stress and the self-report. While this may be true, our findings show that the correlation strengthens as we reduce the window of analysis prior to the micro-EMA, suggesting a potential recall bias. This could also be another reason why self-reports have generated marginal accuracy for longer window sizes.

7 CONCLUSION

Our proposed framework for choosing an optimal micro-EMA question can enable researchers to identify the least burdensome question to be asked in the real world, prior to deployment of the question. By enabling more frequent but less burdensome responses from participants, this framework will enable us to deploy a system that can identify smaller or more subtle changes in physiology to better capture the subjective stress response.

Using our framework, we identified that BinaryStress (r = 0.66), LikertStress (r = 0.63), PSS-Overcome (r = 0.48), and WorriedStress (r = 0.54) (Figure 7) are highly correlated with intended stress, while BinaryStress (kappa = 0.32 and cluster sep. = 0.67), LikertStress (kappa = 0.34 and cluster sep. = 0.68), and WorriedStress (kappa = 0.34 and cluster sep. = 0.69) align most with clustered physiology. We also identified highly predictive features and determined that the most predictive micro-EMA was WorriedStress as it was most accurate when predicting both stress and non-stress events. We provide a framework for individuals to attempt to identify which micro-EMA to use prior to deployment. Future studies should aim to test other feature-selection and clustering algorithms with a greater number of participants in real-world settings.

ACKNOWLEDGMENTS

This work was supported by Robert H. & Ann Lurie Children's Hospital and Stanley Manne Research Institute Perinatal Origins of Disease Strategic Research Initiative and Northwestern University's Institute for Innovations in Developmental Sciences. We would also like to acknowledge support by the National Institute of Diabetes and Digestive and Kidney Diseases under award number K25DK113242. We also gratefully acknowledge contributions of colleagues and collaborators including Drs. William Grobman, Sheila Krogh-Jespersen, Amelie Petitclerc, Karen Mestan, Aaron Hamvas, Mr. Michael Brooks, Ms. Samanvitha Sundar, and Ms. Jayalakshmi Jain.

REFERENCES

[1] Rawan Alharbi, Angela Pfammatter, Bonnie Spring, and Nabil Alshurafa. 2017. WillSense: Adherence Barriers for Passive Sensing Systems That Track Eating Behavior. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 2329–2336.

[2] American Psychological Association et al. 2012. Stress in America: Our health at risk. *Washington DC, American Psychological Association* (2012).

[3] Daniel Bersak, Gary McDarby, Ned Augenblick, Phil McDarby, Daragh McDonnell, Brian McDonald, and Rahul Karkun. 2001. Intelligent biofeedback using an immersive competitive environment. Paper at the Designing Ubiquitous Computing Games Workshop at UbiComp.

[4] Anne-Marie Brouwer and Maarten A Hogervorst. 2014. A new paradigm to induce mental stress: the Sing-a-Song Stress Test (SSST). *Frontiers in neuroscience* 8 (2014), 224.

[5] Jan K Buitelaar, Anja C Huizink, Edu J Mulder, Pascalle G Robles de Medina, and Gerard HA Visser. 2003. Prenatal stress and cognitive development and temperament in infants. *Neurobiology of aging* 24 (2003), S53–S60.

[6] Keng-hao Chang, Drew Fisher, and John Canny. 2011. Ammon: A speech analysis library for analyzing affect, stress, and mental health on mobile phones. *Proceedings of PhoneSense* 2011 (2011).

[7] Jongyoon Choi, Beena 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.

[8] Joseph Ciarrochi, Frank P Deane, and Stephen Anderson. 2002. Emotional intelligence moderates the relationship between stress and mental health. *Personality and individual differences* 32, 2 (2002), 197–209.

[9] C. A. Clark, K. A. Espy, and L. Wakschlag. 2016. Developmental pathways from prenatal tobacco and stress exposure to behavioral disinhibition. *Neurotoxicol Teratol* 53 (2016), 64–74.

[10] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.

[11] Sheldon Cohen, Tom Kamarck, and Robin Mermelstein. 1983. A global measure of perceived stress. *Journal of health and social behavior* (1983), 385–396.

[12] Peter E Comalli Jr, Seymour Wapner, and Heinz Werner. 1962. Interference effects of Stroop color-word test in childhood, adulthood, and aging. *The Journal of genetic psychology* 100, 1 (1962), 47–53.

[13] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20, 3 (1995), 273–297.

[14] Elysia P Davis and Curt A Sandman. 2010. The timing of prenatal exposure to maternal cortisol and psychosocial stress is associated with human infant cognitive development. *Child development* 81, 1 (2010), 131–148.

[15] Tamara Denning, Zakariya Dehlawi, and Tadayoshi Kohno. 2014. In situ with bystanders of augmented reality glasses: Perspectives on recording and privacy-mediating technologies. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2377–2386.

[16] Sally S Dickerson and Margaret E Kemeny. 2004. Acute stressors and cortisol responses: a theoretical integration and synthesis of laboratory research. *Psychological bulletin* 130, 3 (2004), 355.

[17] David F Dinges, Sundara Venkataraman, Eleanor L McGlinchey, and Dimitris N Metaxas. 2007. Monitoring of facial stress during space flight: Optical computer recognition combining discriminative and generative methods. *Acta Astronautica* 60, 4-7 (2007), 341–350.

[18] Hilary Dobson and RF Smith. 2000. What is stress, and how does it affect reproduction? *Animal reproduction science* 60 (2000), 743–752.

[19] Nancy Dole, David A Savitz, Irva Hertz-Picciotto, Anna Maria Siega-Riz, Michael J McMahon, and Pierre Buekens. 2003. Maternal stress and preterm birth. *American journal of epidemiology* 157, 1 (2003), 14–24.

[20] Stewart I Donaldson and Elisa J Grant-Vallone. 2002. Understanding self-report bias in organizational behavior research. *Journal of business and Psychology* 17, 2 (2002), 245–260.

[21] 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 *Pervasive Computing and Communications Workshops (PerCom Workshops), 2017 IEEE International Conference on*. IEEE, 673–678.

[22] S. Entringer, C. Buss, and P. D. Wadhwa. 2010. Prenatal stress and developmental programming of human health and disease risk: concepts and integration of empirical findings. *Curr Opin Endocrinol Diabetes Obes* 17, 6 (Dec 2010), 507–516.

[23] Yoav Freund and Robert E Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences* 55, 1 (1997), 119–139.

[24] Hua Gao, Anil Yüce, and Jean-Philippe Thiran. 2014. Detecting emotional stress from facial expressions for driving safety. In *Image Processing (ICIP), 2014 IEEE International Conference on*. IEEE, 5961–5965.

[25] Mark Andrew Hall. 1999. Correlation-based feature selection for machine learning. (1999).

[26] Kristin E Heron and Joshua M Smyth. 2010. Ecological momentary interventions: incorporating mobile technology into psychosocial and health behaviour treatments. *British journal of health psychology* 15, 1 (2010), 1–39.

[27] Steve Hodges, Lyndsay Williams, Emma Berry, Shahram Izadi, James Srinivasan, Alex Butler, Gavin Smyth, Narinder Kapur, and Ken Wood. 2006. SenseCam: A retrospective memory aid. In *International Conference on Ubiquitous Computing*. Springer, 177–193.

[28] 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 2015 ACM international joint conference on pervasive and ubiquitous computing*. ACM, 493–504.

[29] Tâm Huynh and Bernt Schiele. 2005. Analyzing features for activity recognition. In *Proceedings of the 2005 joint conference on Smart objects and ambient intelligence: innovative context-aware services: usages and technologies*. ACM, 159–163.

[30] Giovanni Iachello, Khai N Truong, Gregory D Abowd, Gillian R Hayes, and Molly Stevens. 2006. Prototyping and sampling experience to evaluate ubiquitous computing privacy in the real world. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM, 1009–1018.

[31] Stephen Intille, Caitlin Haynes, Dharam Maniar, Aditya Ponnada, and Justin Manjourides. 2016. μEMA: Microinteraction-based ecological momentary assessment (EMA) using a smartwatch. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 1124–1128.

[32] Megan M Kelly, Audrey R Tyrka, George M Anderson, Lawrence H Price, and Linda L Carpenter. 2008. Sex differences in emotional and physiological responses to the Trier Social Stress Test. *Journal of behavior therapy and experimental psychiatry* 39, 1 (2008), 87–98.

[33] Clemens Kirschbaum, Karl-Martin Pirke, and Dirk H Hellhammer. 1993. The âĂŸTrier Social Stress TestâĂŹ–a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology* 28, 1-2 (1993), 76–81.

[34] Anne Kouvonen, Mika Kivimäki, Marianna Virtanen, Jaana Pentti, and Jussi Vahtera. 2005. Work stress, smoking status, and smoking intensity: an observational study of 46 190 employees. *Journal of Epidemiology & Community Health* 59, 1 (2005), 63–69.

[35] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174.

[36] Hong Lu, Denise Frauendorfer, Mashfiqui Rabbi, Marianne Schmid Mast, Gokul T Chittaranjan, Andrew T Campbell, Daniel Gatica-Perez, and Tanzeem Choudhury. 2012. Stresssense: Detecting stress in unconstrained acoustic environments using smartphones. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. ACM, 351–360.

[37] Suena H Massey, Amalia E Hatcher, Caron A C Clark, James L Burns, Daniel S Pine, Andrew D Skol, Daniel K Mroczek, Kimberly A Espy, David Goldman, Edwin Cook, and Lauren S Wakschlag. 2017. Does MAOA increase susceptibility to prenatal stress in young children? *Neurotoxicology and teratology* 61 (May 2017), 82âĂŤ91. https://doi.org/10.1016/j.ntt.2017.01.005

[38] P Melillo, C Formisano, U Bracale, and L Pecchia. 2013. Classification tree for real-life stress detection using linear Heart Rate Variability analysis. Case study: students under stress due to university examination. In *World Congress on Medical Physics and Biomedical Engineering May 26-31, 2012, Beijing, China*. Springer, 477–480.

[39] D Out, S Pieper, MJ Bakermans-Kranenburg, and MH van Van IJzendoorn. 2010. Physiological reactivity to infant crying: a behavioral genetic study. *Genes, Brain and Behavior* 9, 8 (2010), 868–876.

[40] Jiapu Pan and Willis J Tompkins. 1985. A real-time QRS detection algorithm. *IEEE transactions on biomedical engineering* 3 (1985), 230–236.

[41] Carmine M Pariante and Stafford L Lightman. 2008. The HPA axis in major depression: classical theories and new developments. *Trends in neurosciences* 31, 9 (2008), 464–468.

[42] Robert S Peirce, Michael R Frone, Marcia Russell, and M Lynne Cooper. 1996. Financial stress, social support, and alcohol involvement: A longitudinal test of the buffering hypothesis in a general population survey. *Health Psychology* 15, 1 (1996), 38.

[43] Kurt Plarre, Andrew Raij, Syed Monowar Hossain, Amin Ahsan Ali, Motohiro Nakajima, Mustafa Al'absi, Emre Ertin, Thomas Kamarck, Santosh Kumar, Marcia Scott, et al. 2011. Continuous inference of psychological stress from sensory measurements collected in the natural environment. In *Information Processing in Sensor Networks (IPSN), 2011 10th International Conference on*. IEEE, 97–108.

[44] Aditya Ponnada, Caitlin Haynes, Dharam Maniar, Justin Manjourides, and Stephen Intille. 2017. Microinteraction Ecological Momentary Assessment Response Rates: Effect of Microinteractions or the Smartwatch? *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 92.

[45] Maria Razzoli, Carolyn Pearson, Scott Crow, and Alessandro Bartolomucci. 2017. Stress, overeating, and obesity: Insights from human studies and preclinical models. *Neuroscience & Biobehavioral Reviews* 76 (2017), 154–162.

[46] Jean E Rhodes and Leonard A Jason. 1990. A social stress model of substance abuse. *Journal of consulting and clinical psychology* 58, 4 (1990), 395.

[47] Akane Sano and Rosalind W Picard. 2013. Stress recognition using wearable sensors and mobile phones. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, 671–676.

[48] Christine Dunkel Schetter and Lynlee Tanner. 2012. Anxiety, depression and stress in pregnancy: implications for mothers, children, research, and practice. *Current opinion in psychiatry* 25, 2 (2012), 141.

[49] Lars Schwabe, Leila Haddad, and Hartmut Schachinger. 2008. HPA axis activation by a socially evaluated cold-pressor test. *Psychoneuroendocrinology* 33, 6 (2008), 890–895.

[50] Suzanne C Segerstrom and Gregory E Miller. 2004. Psychological stress and the human immune system: a meta-analytic study of 30 years of inquiry. *Psychological bulletin* 130, 4 (2004), 601.

[51] Arthur A Stone and Kelly D Brownell. 1994. The stress-eating paradox: multiple daily measurements in adult males and females. *Psychology and Health* 9, 6 (1994), 425–436.

[52] Catherine M Stoney, Mary C Davis, and Karen A Matthews. 1987. Sex differences in physiological responses to stress and in coronary heart disease: a causal link? *Psychophysiology* 24, 2 (1987), 127–131.
[53] Feng-Tso Sun, Cynthia Kuo, Heng-Tze Cheng, Senaka Buthpitiya, Patricia Collins, and Martin Griss. 2010. Activity-aware mental stress detection using physiological sensors. In *International Conference on Mobile Computing, Applications, and Services*. Springer, 282–301.
[54] Constantine Tsigos, Ioannis Kyrou, Eva Kassi, and George P Chrousos. 2016. Stress, endocrine physiology and pathophysiology. (2016).
[55] Bea RH Van den Bergh, Eduard JH Mulder, Maarten Mennes, and Vivette Glover. 2005. Antenatal maternal anxiety and stress and the neurobehavioural development of the fetus and child: links and possible mechanisms. A review. *Neuroscience & Biobehavioral Reviews* 29, 2 (2005), 237–258.
[56] Ruud Van den Bos, Marlies Harteveld, and Hein Stoop. 2009. Stress and decision-making in humans: performance is related to cortisol reactivity, albeit differently in men and women. *Psychoneuroendocrinology* 34, 10 (2009), 1449–1458.
[57] Joe H Ward Jr. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association* 58, 301 (1963), 236–244.
[58] Lida Zhang, Zach King, Begum Egilmez, Jonathan Reeder, Roozbeh Ghaffari, John Rogers, Kristen Rosen, Michael Bass, Judith Moskowitz, Darius Tandon, Lauren Wakschlag, and Nabil Alshurafa. 2018. Measuring Fine-grained Heart-Rate using a Flexible Wearable Sensor in the Presence of Noise. In *IEEE Conference on Biomedical and Health Informatics (BHI) and the IEEE Conference on Body Sensor Networks (BSN)*.