Evaluating the Reproducibility of Physiological Stress Detection Models

VARUN MISHRA, Dartmouth College
SOUGATA SEN, Northwestern University
GRACE CHEN, Washington University in St. Louis
TIAN HAO, IBM T.J. Watson Research Center
JEFFREY ROGERS, IBM T.J. Watson Research Center
CHING-HUA CHEN, IBM T.J. Watson Research Center
DAVID KOTZ, Dartmouth College

Recent advances in wearable sensor technologies have led to a variety of approaches for detecting physiological stress. Even with over a decade of research in the domain, there still exist many significant challenges, including a near-total lack of reproducibility across studies. Researchers often use some physiological sensors (custom-made or off-the-shelf), conduct a study to collect data, and build machine-learning models to detect stress. There is little effort to test the applicability of the model with similar physiological data collected from different devices, or the efficacy of the model on data collected from different studies, populations, or demographics.

This paper takes the first step towards testing reproducibility and validity of methods and machine-learning models for stress detection. To this end, we analyzed data from 90 participants, from four independent controlled studies, using two different types of sensors, with different study protocols and research goals. We started by evaluating the performance of models built using data from one study and tested on data from other studies. Next, we evaluated new methods to improve the performance of stress-detection models and found that our methods led to a consistent increase in performance across all studies, irrespective of the device type, sensor type, or the type of stressor. Finally, we developed and evaluated a clustering approach to determine the *stressed/not-stressed* classification when applying models on data from different studies, and found that our approach performed better than selecting a threshold based on training data. This paper's thorough exploration of reproducibility in a controlled environment provides a critical foundation for deeper study of such methods, and is a prerequisite for tackling reproducibility in free-living conditions.

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

Additional Key Words and Phrases: Stress detection, mobile health (mHealth), wearable sensing, mental health

ACM Reference Format:
Varun Mishra, Sougata Sen, Grace Chen, Tian Hao, Jeffrey Rogers, Ching-Hua Chen, and David Kotz. 2020. Evaluating the Reproducibility of Physiological Stress Detection Models. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 4, Article 147 (December 2020), 29 pages. https://doi.org/10.1145/3432220

Authors' addresses: Varun Mishra, Dartmouth College, varun@cs.dartmouth.edu; Sougata Sen, Northwestern University; Grace Chen, Washington University in St. Louis; Tian Hao, IBM T.J. Watson Research Center; Jeffrey Rogers, IBM T.J. Watson Research Center; Ching-Hua Chen, IBM T.J. Watson Research Center; David Kotz, Dartmouth College.

2474-9567/2020/12-ART147 $15.00
https://doi.org/10.1145/3432220

1 INTRODUCTION

Stress is a dynamic process that reflects the brain's response to internal and external factors: characteristics of a person and her circumstances, as well as the interactions between them [8]. Short-term stress, also known as *acute stress*, can be a positive force by motivating a person to perform well, e.g., before an exam or a job interview, thus leading to positive outcomes. Prolonged periods or frequent occurrence of acute stress may lead to *episodic acute stress*, causing emotional (anger, irritability, anxiety or depression), cognitive (compromised attention/concentration, compromised processing speed, compromised learning and memory) and physical distress (high blood pressure, headaches, digestive problems) [1, 12, 44, 57].

Left untreated, long-term exposure to stress can lead to *chronic stress*, which may significantly and often irreversibly damage a person's physical and mental health. Chronic stress can have a domino effect on other mental and behavioral outcomes, e.g., smoking, drug use, and depression. In fact, the 2019 "Stress in America" survey conducted by the American Psychology Association (APA) found that more than three-quarters of adults reported physical or emotional symptoms of stress, such as headache, feeling tired or changes in sleeping habits; nearly half of adults laid awake at night because of stress in the month prior to the survey; and nearly 3 in 5 adults said they could have used more emotional support in the past year [2]. Hence, stress can be thought of as a fundamental condition, proper management of which can lead to an improvement and balance in the physical, mental, emotional, and behavioral health of an individual.

The current standard method for measuring stress is either via self-reports, like the Perceived Stress Scale (PSS) [14], or by cortisol level measurement [36]. These methods, however, provide momentary insights and cannot be used to monitor or measure stress continuously over an extended period of time, and require a person to go "out-of-their-way", to a clinician or psychologist. Fortunately, recent improvements in wearable sensors and sensing capabilities have enabled researchers to investigate the potential of continuous and passive detection and monitoring of stress in controlled, semi-controlled, and free-living conditions [27, 34, 35, 38–40, 46, 47, 52, 61], leading to over a decade's worth of research and effort in detecting stress using physiological signals.

Despite extensive research, the community is still a long way from producing a validated stress-detection model to deliver interventions and help manage stress. While there are a variety of technical and implementational challenges, there remain several fundamental challenges:

- *Lack of a universally-accepted definition of "stress"*: different people may have a different physiological response to the same "stressor", or report different perceptions of their reaction to stress; in fact, different stressors might result in different physiological responses from the same person.
- *Lack of a "stress signal"*: signals like Electrocardiography (ECG), respiration, or Electro-dermal Activity (EDA) simply capture the physiological response caused by stress. Other aspects of a person's life can cause a similar physiological response, e.g., physical activity, drinking coffee, or engagement in the classroom, thus confounding a stress-detection model.
- *Lack of reproducibility*: researchers often create (or use) custom hardware, conduct a study, and build models. There is little effort to test the applicability of the model with similar physiological data collected from different devices, or the efficacy of the model on data collected from different studies, populations, or demographics.

In this paper, we delve deeper into the third challenge listed above – the *lack of reproducibility*. In the past, researchers have used a variety of different devices and sensors to record physiological signals; for example, Plarre et al., Hovsepian et al. and Sarker et al., used the custom-made AutoSense system for detecting stress in lab and free-living conditions [19, 38, 52, 61], Gjoreski et al. used the commercially available Empatica E4 device for detecting stress in the lab and free-living conditions [27, 28], similarly, Mishra et al. used a commercially available commodity sensor, Polar H7, along with a custom-made EDA sensor, to detect stress in the lab and free-living conditions [46, 47]. While there is an overlap in the physiological signals measured in the different studies, there has been no exploration of whether the machine-learning models built using data collected from one device can apply to the same physiological signal collected from a different device type.

Further, in most of the prior works, the researchers build stress-detection models using data collected from participants belonging to a particular demographic, in a specific setting, using a particular device. Such an approach makes it difficult to evaluate the efficacy of the models applied to a broader population, or to compare with the methods used in other studies. The model reproducibility is either assumed or left as future work, without any quantifiable evidence.

In this work, we take a step towards testing reproducibility of machine-learning models for stress detection. We look at data collected from 90 participants, in four independently conducted studies, with different participant demographics (including high-school students, university students, and corporate employees), using two different types of sensors, with different study protocols and research goals. In all of the studies, the data was collected in a controlled setting, so our work in testing reproducibility and validity of models is limited to controlled settings. However, we argue that testing reproducibility in a controlled environment is a foundational step before reproducibility in free-living conditions can be ensured.

While our initial goal was to test reproducibility, we further discuss and evaluate new methods and techniques to improve the performance of detecting stress, and test the reproducibility of those methods across the four studies. In prior work, the basic “framework” for detecting stress has largely been the same: (1) clean the signal, (2) normalize the data, (3) break the signal into fixed-time windows, (4) compute features over individual windows, (5) train classifiers with the computed features, and (6) measure performance on a test dataset or with some form of cross-validation. Most of the prior work has focused on optimizing and improving the various steps, by testing different normalization techniques, different cleaning techniques, different features, different size time windows, and different machine-learning models. None of these prior works, however, account for the temporal dynamics of stress across time windows, and thus miss out on potentially important information that could improve detection performance. Mishra et al. proposed a two-layered approach, in which they account for the stress in the previous minute while making an inference about the stress in the current minute, and observed a significant boost in performance, as compared to the standard single-model classification [47]. We evaluate the effectiveness of the modeling strategy proposed by Mishra et al., across the different studies, and compare it with other approaches to leverage the temporal dynamics of stress.

Further, we propose a new method to determine the *stressed/not-stressed* classification threshold. In the past, researchers have either used direct binary classifications, or a probabilistic threshold to determine what is classified as “stress”. To maximize detection performance, this threshold is either optimized to the study or to the individual. In this paper, we test the effectiveness of a threshold learned from one study when evaluating the model with data collected from another study, and show that such a “fixed” threshold does not work across different studies. We thus propose an unsupervised method to determine the classification threshold that does not require “hard-wired” thresholds and adjusts the threshold according to measured stress probabilities.

Finally, we do recognize that our work is focused on controlled studies only, and does not account for the challenges and complexities in free-living conditions. We discuss how to deploy models trained in the lab studies to real-life situations, however, and share insights on how to improve performance.

1.1 Defining Repeatability, Replicability, and Reproducibility

It is important to note that there are no “fixed” definitions for the terms “repeatability," “replicability," and “reproducibility" [54]. As highlighted by Hans Plesser, researchers have used contradictory definitions for these terms for many years [54]. For our context of stress detection, we take inspiration from the definitions proposed by Crook et al. [15], and the ones adopted by ACM in 2016 [20], to define the terms as follows.

Repeatability (i.e., same team, same experimental setup): when the same team of researchers can get the same/similar results (as the original study) in later studies using models or methods from the original study.[1]

Replicability (i.e., different team, same experimental setup): when a different team of researchers conducts an independent study by following the experimental setup of a previous study and can achieve similar results (as the original study) by using the models or methods from the original study.

Reproducibility (i.e., different experimental setup): when a team of researchers use models or methods from a previous study and achieve similar results from an independent study with a different experimental setup.

Under these definitions, a machine-learning model built on Study X can be considered *repeatable* if it achieves similar results when tested on data from Study Y (where Studies X and Y were conducted by the same researchers and had the same research protocol). A model from Study X would be *reproducible* if it achieves similar results when tested on data from Study Z (where Studies X and Z followed different research protocols). In similar vein, methods for detecting stress from Study X would be *reproducible* if their application on train/test data collected from Study Z results in similar results or outcomes.

It is important to note that reproducibility is different from *generalizability*. A model trained on Study X could be reproducible on data from Study Z, and still not be generalizable. We argue that reproducibility is the first step towards generalizability. To be considered *generalizable*, models should be consistently reproducible across a large number of studies with varying characteristics (sensors, devices, and demographics). Testing for generalizability would involve several years of reproducible research and is clearly beyond the scope of this work.

2 RELATED WORK

Many prior works have aimed to *detect stress*. Although there have been efforts to use 'contact-less' strategies for detecting stress, like using smartphone data (e.g., call/SMS logs, app usage, or motion [23, 23, 68]), or by processing a user's voice [42], or by processing facial expressions [21], or by analyzing keyboard typing behaviors [58], in this paper we focus only on methods that employ wearable devices to measure physiological signals to study stress. While such contact-less methods have some advantages, they cannot provide a fine-grained – minute-by-minute – continuous assessment of stress, and some methods induce privacy risks.

In the domain of physiological stress sensing, prior works have used a variety of wearable sensors, e.g., ECG sensor [28, 38, 39, 47, 52], EDA sensor [11, 17, 28, 35, 47, 66], Respiratory Inductance Plethysmography (RIP) sensor [28, 35, 38, 52, 61], Blood Volume Pulse (BVP) sensor [28], or Electromyogram (EMG) sensor [35]. Some studies employed these sensors alone; others explored combinations of two or more sensors.

Further, these sensors have been used in a variety of conditions: (a) stress induced in a lab situation, where the participants undergo some well-validated stress-inducing task, e.g., mental arithmetic, public speaking, stroop test, startle response test, or cold-pressor test, along with some restful periods to serve as baseline [17, 28, 38, 39, 47]; (b) constrained real-life activities, where researchers could monitor participants' stress levels as they engaged in a particular activity, e.g., driving [35], in a call center [37], or while sleeping [48]; (c) in free-living conditions, where the participants wear the sensors as they continue with their daily lives [28, 38, 39, 47, 52].

For clarity, we summarize the various prior works in Table 1. We report the type of environment/situation(s) where the study was conducted, the type of data collected in those studies, the types of sensors/devices used, the number of participants, and the results obtained by the authors.

From Table 1, we observe that (in almost every case) the machine-learning models and subsequent evaluations have been based on a relatively small number of participants belonging to a narrow set of demographics. Each study resulted in the building and evaluation of some machine-learning model. There has been little effort to evaluate models over a larger population and broader demographic. Hence, it is unrealistic to compare the

[1] Given the natural variability in human-sensing research, *same results* can be construed as *similar results*, i.e., the results are not significantly different.

methods, devices or model performances across different studies. Only a few of the prior works have looked into the validation of their model. Hovsepian et al. built a stress-detection model using data from a lab study with 21 participants [38]. They further evaluated the *repeatability* of their model on another in-lab study with 26 participants. Both in-lab studies followed the same study protocol, used the same sensor-suite with ECG and respiration sensors (AutoSense [19]), and included participants belonging to the same demographic (university students). They found that the model performance with data from the new study was similar to the cross-validation performance in the initial study (i.e., the model was repeatable). Hovsepian et al. also conducted a field study with 20 participants and adapted their in-lab model to detect stress in the free-living conditions and developed the *cStress* model.

Further, Sarker et al. used the cStress model to detect stress among the 38 polydrug users in their independently conducted field study [61]. The authors used the same sensor-suite as Hovsepian et al. and found that they were able to detect self-reported stress with an F1-score of 0.717, which was similar to the F1-score achieved by Hovsepian et al. in their field study. However, in their work, Sarker et al. followed extensive methods to filter the data and impute missing data before computing features to be used by the cStress model. Even the model was modified to provide outputs at a much faster frequency than the original model. Given the significant difference in the methods used for feature calculation before the model was applied, it is unknown what role the different methods had in the stress-detection performance. Thus, the reproducibility of the model itself cannot be evaluated.

Hence, there is a clear need to evaluate how models built using data from one study translate to data collected from a different study with different participant demographics and research protocols. In this paper we take a step in this direction. We evaluate the reproducibility of stress-detection models across four independently conducted studies. The studies were conducted using two different types of devices, with different research protocols, and include data from 90 participants belonging to different demographics.

Another key observation from Table 1 is that stress-detection models based on physiological sensing work reasonably well in lab-based or constrained situations, but they perform poorly when deployed in free-living conditions – a pattern consistent across studies. One key reason is that the wearable sensors used in these studies do not really measure *stress*, they simply measure the body's physiological response to stress. As noted above, free-living conditions can cause physiological responses that confound stress-detection models. Several prior works try to account for physical activity in free-living conditions to remove such confounds, which does help improve the model performance [28, 38, 47, 61]. Accounting for physical activity alone, however, might not be sufficient. Researchers use self-reports to collect ground-truth for detecting stress in free-living conditions. Self-reports capture a person's *perception* of stress, which could be quite different from a physiological measure of stress, or the actual underlying condition of stress. It is possible that a person does not perceive a scenario as stressful and even though they are physically inactive, their physiological response could show that they are stressed. In some preliminary results, Mishra et al. show that taking the high-level context of an user into consideration (e.g., working, sleeping, eating, in a meeting, or resting) could help improve detection of physiological stress [45].

3 THE DATA

In this paper we evaluate data collected from four different and independently conducted studies, and a total of 100 participants (in this work we used data from 90 participants). All the studies were conducted between 2017 and 2019. We conducted one of the studies and data from the other three were graciously shared by their respective authors. In this section, we discuss the different studies, the respective study protocols, and the type of devices used. All the studies were approved by their respective Institutional Review Board (IRB) or Ethics Committee.

Table 1. Summary of Related Work

	Setting	**# of Participants**	**Data used**	**Devices used**	**Results**
Choi et al. [11]	Lab	10	HRV, RIP, EDA and EMG	Custom chest–strapped sensor suite	Binary classification with 81% accuracy
Hernandez et al. [37]	Call Center	9	EDA	Affectiva Q Sensor [30]	Personalized model: 78.03% accuracy. Generalized model: 73.41% accuracy
Egilmez et al. [17]	Lab	9	Heart-rate, EDA, Gyroscope	Custom EDA sensor, with LG Smartwatch	Binary classification with F1 score of 0.888
Sano et al. [60]	Field	18	EDA and Smartphone usage	Affectiva Q sensor [30], and smartphones	Binary classification using 10-Fold cross-validation resulted in 75% accuracy.
Plarre et al. [52]	Lab, Field	21	ECG and RIP	AutoSense sensor suite [19]	Lab: Binary classification of stress with 90.17% accuracy Field: High correlation (r=0.71) with self-reports
Hovsepian et al. [38]	Lab, Field	Lab Train: 21 Lab Test: 26 Field: 20	ECG and RIP	AutoSense sensor suite [19]	Binary classification: Lab Train LOSO CV F1 score: 0.81 Lab Test F1 score: 0.9 Field self-report prediction F1 score: 0.72
Sarkar et al. [61]	Field	38	ECG and RIP	AutoSense sensor suite [19]	Field self-report prediction F1 score of 0.717 by using the cStress model.
Sun et al. [66]	Lab	20	ECG and EDA	Custom chest- and wrist-based sensor suite	Binary classification accuracy by 10-fold CV was 92.4%. Accuracy for cross-subject classification was 80.9%.
Gjoreski et al. [27]	Lab, Field	Lab: 21 Field: 5	BVP, EDA, HRV, Skin Temperature, Accelerometer	Empatica E3 [22] and E4 [18]	Lab: LOSO accuracy of 72% when classifying between no stress, low stress and high stress. Field: Binary classification for detecting stress with F1 score of 0.81
King et al. [39]	Lab, Field	Lab: 18 Field: 17	ECG	BioStampRC [43]	Binary classification of stress: Lab LOSO CV F1 score: 0.70 Field Accuracy: 62%
Mishra et al. [47]	Lab, Field	26	ECG and EDA	Polar H7 [55] and custom EDA sensor [56]	Binary classification of stress: Lab LOSO CV F1 score: ECG only: 0.87, ECG+EDA: 0.94 Field LOSO CV F1 score: ECG only: 0.66, ECG+EDA: 0.73

3.1 Study 1 (S1)

This study was conducted with 26 participants by Mishra et al., using the Polar H7 heart-rate monitor with the Amulet device [6, 46, 55]. All the participants were students at a US university, and included a mix of undergraduate and graduate students.

In this study, the participants went through three different stress-inducing tasks: (a) mental arithmetic task, in which the participants were asked to count backwards in steps of 7; (b) a startle response test, in which participants faced away from the research staff with their eyes closed, and the staff then made a loud sound at several random times to startle the participants; (c) the cold pressor task, in which the participants were asked

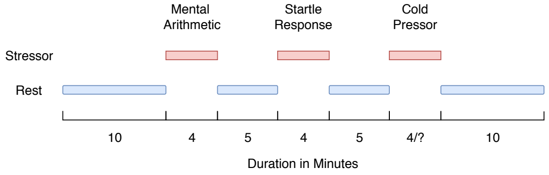

Fig. 1. Lab protocol timeline for Studies 1 and 2. In Study 2, however, the order of the stressors was randomized.

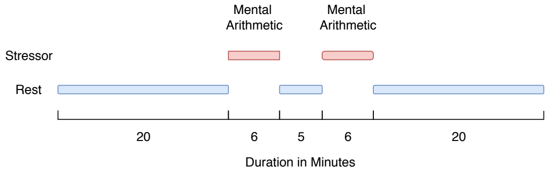

Fig. 2. Lab protocol timeline for Study 3.

to submerge their hand in a bucket of ice water for four minutes or as long as tolerable. A timeline of the lab protocol is shown in Figure 1.

3.2 Study 2 (S2)

We conducted this study following a protocol similar to the one used by Mishra et al. [46], with 13 participants at a US university. We used the Polar H10 heart-rate monitor (an updated version of the Polar H7). The participants underwent the same three stress-inducing tasks, as in $S1$: (a) mental arithmetic task, (b) startle response test, and (c) the cold pressor task. Unlike $S1$, however, we randomized the sequence of the tasks, i.e., each participant underwent a different sequence of stressors. The timeline of the lab protocol is shown in Figure 1.

3.3 Study 3 (S3)

This study was conducted by Hao et al., with 19 participants, all of whom were employees at a corporate organization [32, 33]. The participants were asked to wear an Empatica E4 wrist sensor [18] as they underwent some mental arithmetic tasks. The lab session started with a 20-minute relaxation period, followed by two 6-minute periods of mental arithmetic tasks with a 5-minute break in between. For each of the stressful periods, the participants were verbally asked non-trivial mental arithmetic tasks (e.g., 2010 − 37 =?), every 10 seconds. A visual representation of the stages in the lab session is shown in Figure 2.

3.4 Study 4 (S4)

This study was conducted by Chen et al. with 42 participants from a US high school [10]. The goal of this study was to understand the effect of listening to music on physiological stress. In this study, Empatica E4 wrist devices were used to collect physiological data. This study protocol included only one 5-minute-long mental arithmetic stressor, in which the participants were asked to count backwards in steps of 13.

Given the goal of the study, the participants were randomly divided into four groups: the control group (C) did not listen to music, one experimental group (M1) listened to music for 5 minutes before the stressor, the second experimental group (M2) listened to music for 5 minutes after the stressor was applied, and the final experimental group (M3) listened to music for 10 minutes after the stressor was applied. A timeline of the lab protocol split by the different groups is shown in Figure 3.

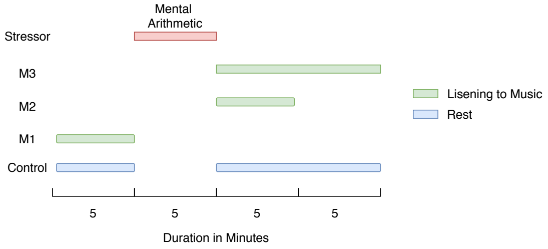

Fig. 3. Lab protocol timeline for Study 4.

3.5 Data Summary

The four datasets mentioned above (a) use two different types of devices for collecting physiological signals: a commodity chest-based heart-rate monitor (Polar H7/H10) and a research-grade wrist-based wearable (Empatica E4), (b) target different demographics, (c) follow different protocols, and (d) were collected at different times, thus making it a heterogeneous collection of data. While some aspects vary, e.g., in *S*4 participants were listening to music at different times, or in *S*1 & *S*2, participants underwent different types of stress tests, all the datasets have a well-defined baseline (rest) period and at least one stress-induction period, thus enabling us to effectively build stress-detection models and compare performance across studies.

In *S4*, according to the study protocol designed by Chen et al., the M1 group (10 participants) listened to music during the baseline rest period [10]. In their analysis, the authors found that the average detected stress probability during baseline rest period for the M1 group was significantly higher than the C, M2, and M3 groups. Since listening to music caused some physiological arousal for the M1 group, we decided to discard the M1 group from our analyses. Ultimately, we included a total of 90 participants in our analysis across the four studies.

We used the baseline (rest) and stress-induction periods as the 'ground truth' to evaluate the performance of the models by categorizing them as *not-stressed* (class=0) or *stressed* (class=1), respectively.

4 METHOD

In this section, we discuss our choice of physiological signals, methods for processing data, feature selection, classification tasks and evaluation criteria.

4.1 Choice of Physiological Signals

Across the four studies, we considered data from the Polar H7, Polar H10, and the Empatica E4. The Polar H7 and H10 are chest-based heart-rate monitors. They do not provide raw ECG data; instead they directly provide the heart-rate and R-R interval data. Empatica E4 is a wrist-based sensor and relies on BVP to calculate heart-rate and R-R interval data. Empatica E4 provides both the raw BVP data and the processed heart-rate and Inter-Beat Interval (IBI) data.[2] Although many approaches exist for processing raw BVP data, we use only the heart-rate and R-R interval (IBI) data provided by the E4. This approach can enable other researchers to directly use our methods. Furthermore, this approach is consistent with our use of the data from the Polar H7/H10; in each case we use the processed data produced from a commercially engineered device. We were thus able to use the same downstream processing pipeline for both device types.

Further, the E4 also measures EDA[3] and skin-temperature data. Some prior work has used IBI or EDA sensors – sometimes with other sensor streams, like EMG, RIP, or skin temperature – to detect stress [17, 28, 35, 37, 38, 47, 52]. While the Empatica provides IBI, EDA, and skin-temperature data, we only use the former two sensor streams, and ignore the skin temperature data because prior work has found skin temperature from the Empatica E4 to be a poor indicator of stress [28, 62].

In summary, we use two different physiological signals: heart rate (and R-R interval) in all four studies, and EDA in the two studies using Empatica.

4.2 Data Processing

Because we used two types of physiological signals, we followed methods suitable for each signal rather than a common approach for both.

[2] For practical purposes, the R-R interval and IBI are the same thing. IBI represents the time between two beats, and R-R interval represents the time between two consecutive R peaks in an ECG.

[3] The terms EDA and GSR are often used interchangeably. In this paper, we use the term EDA.

4.2.1 Data Cleaning. The goal of this step was to remove clearly erroneous data points. For heart-rate data, we followed the criteria used by Mishra et al. [46]: if the heart-rate value was not within the range [30:220] we discarded the heart-rate and any R-R interval values received during that second [46]. Similarly, for EDA data, we discarded values below $0.01\mu S$ and above $100\mu S$.

4.2.2 Outlier Handling. Although the participants were sitting in a controlled enviroment, sudden hand or body movements could introduce outliers in the data. For heart-rate and R-R interval data, there are two popular strategies for handling outliers: *winsorization* [28, 38] and *trimming* [46]. Mishra et al. found trimming led to better classification results than winsorization, so we adopt that strategy for heart-rate and R-R interval data: we set the threshold at 3 times the Median Absolute Deviation (MAD). Thus, we trimmed any data points outside the $median \pm 3 \times MAD$. We took a slightly different approach for EDA data, however. Prior research has found filtering techniques to be effective at handling outliers or artifacts in EDA signals from the Empatica E4 device [16, 25, 62]. Although there are a variety of filtering methods, we used a median filter with a 5-second window as used by Di Lascio et al. [16]. Median filters help reduce artifacts while preserving the edges of the signal.

4.2.3 Data Normalization. Normalization is a important step to remove participant-specific effects (like mean or standard deviation), and potentially make the same model generalizable to other users, without needing to create personalized models. For heart-rate and R-R data, we used *z*-score normalization as recommended by Mishra et al. [46]. For EDA signals, given that individuals tend to have different natural EDA ranges, we used *min-max* normalization to bring the signals into the range $[0, 1]$.

4.2.4 EDA Decomposition. EDA data requires an additional processing step. There are two main components to the overall EDA signal: *tonic* and *phasic*, and it is common practice to decompose the EDA signal into these components [7, 16, 25, 62]. The tonic component relates to the slowly varying signal and measures physiological arousal. The phasic component represents the faster-changing elements and is characterized by rapid changes and spike-like features [65]. To decompose the EDA signal, we used the *cvxEDA* method by Greco et al., which uses convex optimization to decompose the signal [13, 29].

4.3 Feature Extraction

In the past researchers have commonly used a 60-second window for detecting stress with physiological signals [17, 38, 39, 46]. In fact, Esco et al. demonstrated that, when compared to 10- or 30-second window size, the features computed on a 60-second window size had the highest agreement with the conventional 5-minute window. Hence, for feature extraction, we grouped the data into 60-second sliding windows with 75% overlap, i.e., windows overlapped by 45 seconds.[4] We outline below the features collected from the different signals.

For heart-rate and R-R interval signals, we computed several time-domain features, as listed in Table 2. Prior works have demonstrated that these features show significant differences between stress and non-stress periods, and can be used to detect stress. In our work, we avoided using frequency-domain features for several reasons.

- Prior work has shown that RMSSD (root mean square of successive differences of successive R-R intervals) is a solid measure of vagal tone and parasympathetic activity [41]; much like HF (High Frequency energy). Further, Shaffer et al. show that RMSSD and HF are strongly correlated [63]. Further, RMSSD is easier to compute and is not confounded by breathing.
- While HF represents parasympathetic activity, the role of LF is unclear. Some researchers believe that LF represents sympathetic activity, thus making the LF:HF ratio a representation of the sympatho-vagal balance. Other researchers, however, argue that the LF is not a pure index of sympathetic activity, but instead

[4] Prior works have evaluated different lengths of sliding windows, like overlap [38, 47], 50% overlap [39], 75% overlap [17], and even 92% overlap [61]. During some preliminary evaluations with the $S2$ data, we found that 75% overlap resulted in the highest F1-score, when compared to other overlap periods (25%, 50%, and 92%). Thus we chose 75% for all the evaluations in this paper.

is a non-linear combination of sympathetic and parasympathetic activities [5, 63]. Thus the usefulness of LF as an indicator for detecting stress is not clear. Further, work done by Hovsepian et al. found that LF and LF:HF features had very low importance [38].

- Finally, to reliably calculate LF features, a minimum time window of 2 minutes is required [64]. This would reduce our sample size by half.

Hence, given the lack of a clear benefit of using frequency-domain features, we used just the time-domain features in our work.

For EDA data, we calculated various statistical features as mentioned by Lascio et al. [16]. These features are calculated for the overall EDA signal, as well as the tonic and the phasic components. These features have been shown to capture general arousal of a person [16]. The EDA features are listed in Table 2.

Table 2. Features from Heart-rate, R-R interval, and EDA data.

Signal Type	**Features**	
Heart-rate	mean, median, max, min, standard deviation, kurtosis, skew, slope, 80th percentile, 20th percentile	
R-R interval	mean, median, max, min, standard deviation, kurtosis, skew, slope, 80th percentile, 20th percentile, RMSSD	
EDA	EDA Tonic Phasic	mean, max, min, standard deviation, number of peaks, area under curve (AUC)

4.4 Data Labeling

All three datasets from prior studies (S1, S3, and S4) were divided in two classes: *stressed* and *not-stressed*. The time windows representing the baseline period for all datasets were labeled as *not-stressed* (class 0), since the participants were not undergoing any stress tasks. Any time window during which a participant was experiencing a stressor was labeled as *stressed* (class 1). To be able to effectively compare performances across datasets, we followed the same labeling strategy for the data collected from study $S2$.

Further, Mishra et al. found that there might be some residual physiological stress in the baseline rest period, and observed that considering only the last 4 minutes (i.e., discarding the first 6 minutes of the 10-minute initial rest period) as the *not-stressed* labels led to a significant improvement in classification performance [46]. Based on this empirical evidence, in our work, we decided to exclude some windows at the beginning of the the baseline rest period to remove any residual physiological stress.[5,6]

Further, the $S1$ and $S2$ studies included three different types of stressors: mental arithmetic, startle response, and cold pressor, whereas the $S3$ and $S4$ studies included just the mental arithmetic stressor (twice in $S3$ and once in $S4$). Mishra et al. observed that different stressors might have different physiological responses, evidenced by the fact that the authors were able to distinguish among stressor types with an average F1-score of 0.71 [46]. Hence, to ensure a direct comparison between the four studies, we built models and compared the performance

[5] We excluded first 6 minutes for studies $S1$ to $S3$, and the first minute for study $S4$. Since study $S4$ had a baseline rest duration of only 5 minutes, dropping longer windows from $S4$ would adversely affect the amount of "rest" periods.

[6] Consistent with prior work, we do not include the "intermediate" rest periods (between or after stressors) in our model building or evaluation [46, 47, 52]. We argue that such rest periods will contain some residual physiological arousal of the preceding stressor, and hence labeling them *not-stressed* might not be appropriate. While these "intermediate" periods can potentially provide insights about stress recovery, we focus our current work only on *detection* of stress.

with only the mental arithmetic task(s) as the *stressed* label, and the baseline rest period as the *not-stressed* label. For distinction, we name the subsets $S1_{math}$ and $S2_{math}$, respectively. For completeness, however, we believe it is important to evaluate the performance of the models across the different types of stressors. Thus, we also discuss the performance of the models built using just the mental arithmetic stressor on the $S1_{all}$ and $S2_{all}$ datasets, which include the time windows for all three stressors as *stressed*.

Finally, two of the studies ($S1$ and $S2$) did not have any EDA data. Hence, we decided to split the datasets from $S3$ and $S4$ into two categories, one that included only heart-rate and R-R interval features, and the other with heart-rate, R-R interval, and EDA features. The models built with the combination of heart-rate, R-R interval, and EDA features – henceforth named $S3_{eda}$ and $S4_{eda}$ – could only be evaluated and compared with models from the $S3$ and $S4$ studies; whereas the subsets of $S3$ and $S4$ with just the heart-rate and R-R interval features – named $S3_{hr}$ and $S4_{hr}$ – can be evaluated and compared with models built on $S1$, $S2$, $S3$, and $S4$ studies.

4.5 Machine Learning Models

Researchers have used a variety of machine-learning algorithms to detect stress, e.g., Naive Bayes, KNN, Decision Trees, Support Vector Machines (SVM), Random Forests (RF), Multi-layer Perceptron, AdaBoost, and Logistic Regression; however, several studies have found that SVM and RF perform better than other models [17, 28, 38, 39, 46]. Both SVM and RF tend to limit over-fitting and reduce the bias and variance of the resulting models.

While there is no clear consensus as to which classifier performs best, Mishra et al. noted that SVM performs better for models including just heart-rate and R-R interval features, whereas RF performs better when the models included heart-rate, R-R interval, and EDA features [46]. In our work, we compare performance from both types of model, and choose the better performing model to continue our analysis.

Most prior works in stress detection have used Radial Basis Function (RBF) as the kernel for their SVM models [28, 38, 39, 46]. Based on preliminary cross-validation evaluations with data from study $S2$, we found that RBF SVM performed substantially better than Linear SVM. Hence, we decided to use RBF kernel. It has two hyper-parameters: C and γ, the choice of which can significantly affect the results of the SVM algorithm. The usual approach is to conduct an exhaustive grid-search evaluated using Leave-One-Subject-Out (LOSO) cross-validation to find the values of the hyper-parameters that maximize a performance metric like F1-score. In prior work, Hovsepian et al. reported that setting $C = 724.077$ and $\gamma = 0.022097$ led to the best performance of their model built with Heart Rate Variability (HRV) data [38]. In later work, King et al. did an exhaustive grid search and found that $C = 107$ and $\gamma = 0.001$ led to best performance in their model with HRV data [39], even when compared with the parameters reported by Hovsepian et al.

In our work, we began with the same approach – conducting an exhaustive grid-search to build models for our various datasets – and compared the results with the hyper-parameters reported by King et al. and Hovsepian et al. Although we found different values of the hyper-parameters, we found no significant difference between the performance achieved by our "tuned" hyper-parameters and those reported by King et al. Because we compare multiple models over four different studies, for consistency we set the hyper-parameters for all SVM models to what was reported by King et al.: $C = 107$ and $\gamma = 0.001$.

For each input observation, the SVM and RF models output the probability of the *stressed* class. We used the LibSVM library for building the SVM model [9]. LibSVM uses Platt's scaling to transform the canonical distance between of input observation from the hyper-plane into a conditional probability [53]. For Random Forests, we used the Scikit-learn library, which can output the predicted class probabilities of an input sample by "computing the mean predicted class probabilities of the trees in the forest. The class probability of a single tree is the fraction of samples of the same class in a leaf" [51].

In the next section, we outline our iterative process for building machine-learning models and discuss the results obtained by evaluating each model type.

5 MODEL BUILDING & EVALUATION

In this section we discuss the machine-learning models we built along with the evaluation of those models. We start by discussing study-specific models, which were evaluated by LOSO cross-validation. We then discuss 'cross-study' models – trained using data from one study and evaluated on data from other studies – along with methods to improve the model performance.

5.1 Study-specific Models

In the first step of our analysis, we built models specific to each study and performed LOSO cross-validation within each of the four studies (six models trained using the mental arithmetic stressor). This process established a reference to compare performance of models built using training data from a different study. Ideally, it is desired that the performance of models built using data from a different study remains close to (if not better than) the performance of the LOSO cross-validated model.

As we mention in Section 4.5, to compare the utility of SVM and Random Forests, we built both types of model for each of the datasets and evaluated them through LOSO cross-validation. In total, we built two models for each of the following datasets: $S1_{math}$, $S2_{math}$, $S3_{hr}$, $S3_{eda}$, $S4_{hr}$, and $S4_{eda}$. For completeness we also report the LOSO cross-validation results for the models built on $S1_{math}$ and $S2_{math}$, and evaluated on $S1_{all}$ and $S2_{all}$, respectively. In such cases, the model was trained using only the mental arithmetic stressor, but evaluated on the mental arithmetic, startle response, and cold pressor stressors of the *left-out* user.

Both SVM and RF models output a probability of stress. Prior works by Hovsepian et al. and Mishra et al. used a *threshold*-based approach, such that any instance with a probability of stress greater than a particular threshold was classified as *stressed* [38, 47]. In both of the prior works, the threshold was chosen on the basis of maximizing the F1-score in LOSO cross-validation. We followed the same approach for determining the threshold(s). Based on the threshold, we classify each 60-second window as *stressed* or *not-stressed*. We report the following metrics: *precision*, fraction of instances labeled "positive" that were actually positive; *recall*, the fraction of positive instances correctly labeled as positive; and *F1-score*, the harmonic mean between precision and recall, which is a popular metric in classification problems with one primary class of interest.

Based on our evaluations, we observe that SVM performed better for datasets that only had heart-rate and R-R interval features, whereas Random Forest performed better for datasets that also included EDA features (in addition to heart-rate and R-R interval features), consistent with findings by Mishra et al. [47]. Hence, we only report the results from best-performing models for each data subset in Table 3. We also list the classification threshold for each model.

Table 3. LOSO cross-validation results for the different data subsets.

Metrics	$S1_{math}$ (SVM)	$S1_{all}$ (SVM)	$S2_{math}$ (SVM)	$S2_{all}$ (SVM)	$S3_{hr}$ (SVM)	$S3_{eda}$ (RF)	$S4_{hr}$ (SVM)	$S4_{eda}$ (RF)
Precision	0.87	0.78	0.82	0.79	0.65	0.79	0.63	0.84
Recall	0.86	0.70	0.76	0.72	0.66	0.82	0.83	0.85
F1-score	0.86	0.74	0.79	0.75	0.66	0.80	0.72	0.85
Threshold	0.39	0.21	0.45	0.24	0.51	0.50	0.41	0.49

We observe that the performance and threshold of cross-validation using $S1_{math}$ was quite different than the model built on $S2_{math}$. This is an interesting observation, since both studies had similar protocol and sensors. We believe it is because of the randomized order of the stress induction tasks. Further, when we evaluated models built using just the mental arithmetic stressor from $S1$ and $S2$, on the other stressors in their respective studies, we found that the recall of the models dropped. This suggests that the model was not able to identify all *stressed* periods. It could be because the other stressors (startle response and cold pressor) might not result in a similar physiological response as the mental arithmetic test.

We further observe that the performance of the models built using the heart-rate features from Empatica E4 ($S3_{hr}$ and $S4_{hr}$), perform substantially worse than models built using the Polar H7 and H10 devices, at similar stress detection tasks (mental arithmetic vs. baseline rest). One reason could be that while Polar H7 and H10 sensors are one-lead ECG devices that measure the electrical activity of the heart (and are known to be accurate [26]), Empatica E4 uses a Photoplethysmogram (PPG) sensor to measure blood-volume changes to calculate heart-rate and IBI data, and the Empatica has recently been shown to have errors when compared with an ECG patch [3]. Upon adding the the EDA features, however, there is an improvement in performance.

During our initial evaluations, we made an important observation: the performance of the model changes substantially with the classification threshold. Further, this threshold changes with device type (i.e., data quality) and the data distribution in training and test sets. For example, the threshold (0.39) for $S1_{math}$ is very different than the threshold (0.21) for $S1_{all}$ (which includes all three stressors in the study). As in most prior works, we chose a threshold based on LOSO cross-validation, and it works well because of homogeneity in the data: each participant's data is expected to have a similar class distribution as they all underwent similar study procedures. However, using the same threshold for classification on independent and unseen *test* users with no a priori knowledge of the data distribution is a challenge. Tuning the threshold (or any hyper-parameter) using the *test* set will violate the independence of the data and the results will be optimistically biased.

In the next section, we build models using data from one study and evaluate using data from another study, without making any assumptions about the data distributions of the *test* set. Hence, to evaluate the model performance, we use the Area Under the Receiver Operator Characteristics curve (AUROC) as the performance metric [31]. The Receiver Operator Characteristics (ROC) curve plots the True Positive Rate (TPR) against False Positive Rate (FPR) for different classification thresholds. The AUROC is a metric that summarizes the ROC curve, and represents the ability of the model to distinguish between the two classes. The AUROC ranges from 0.5 to 1.0; an AUROC of 1.0 means the classifier is able to perfectly separate between the two classes, and a score of 0.5 means the classifier is no better than a random guess. Technically, AUROC could also have a value less than 0.5; a value of 0.0 means that the classifier is able to separate between the classes, but it flipped the labels (i.e., marked positive as negative and vice-versa). Also, since AUROC is based on TPR and FPR, it does not vary with changes in class distribution of the test set, unlike precision and recall. Hence, AUROC is widely used to compare performance of binary classifiers [31].

We argue that building a model and choosing a decision (classification) threshold are two separate components. Model building ends with a probability output, and then based on the application domain, a decision threshold may be selected. We argue it is important to build good robust models, before *tuning* or selecting a decision threshold. In the next sections, we discuss how cross-study models perform, how we might be able to improve model performance, and finally how to make a decision about stressed or not-stressed without a pre-defined threshold.

5.2 Cross-study Models

Given the diversity in sensors and stress-inducing tasks in our dataset, we built different models for each of our data subsets, and evaluated the performance on different data subsets containing the same features. Specifically, we built models trained on: $S1_{math}$, $S2_{math}$, $S3_{hr}$, $S3_{eda}$, $S4_{hr}$, and $S4_{eda}$. We trained SVM and Random Forest (RF) models for each data subset. All models were evaluated on each participant in the evaluation set, and generated participant-level AUROC. We present the median scores (along with the Interquartile Range) from the best performing models (SVM or RF) in Table 4. Each cell in Table 4 represents the median AUROC when a model

Table 4. Cross-study Evaluations: Each cell represents the median AUROC for a model built on the training set and tested on the evaluation set. The Interquartile Range (IQR) is shown in square brackets. In case the training and evaluation datasets are from the same study, we report the results of a LOSO cross-validation. The blue cells show results from a SVM model; the yellow cells show results from a Random Forest (RF) model. As data from studies $S1$ and $S2$ did not include any EDA data, models built with the combination of heart-rate, R-R interval, and EDA features – specifically $S3_{eda}$ and $S4_{eda}$ – could only be evaluated and compared with models from the $S3$ and $S4$ studies.

Training Dataset	***Evaluation Dataset***							
	$S1_{all}$	$S1_{math}$	$S2_{all}$	$S2_{math}$	$S3_{hr}$	$S3_{eda}$	$S4_{hr}$	$S4_{eda}$
$S1_{math}$	0.67 [0.48 – 0.77]	0.98 [0.93 – 1.00]	0.62 [0.57 – 0.72]	0.93 [0.81 – 0.97]	0.75 [0.62 – 0.89]		0.80 [0.50 – 0.94]	
$S2_{math}$	0.61 [0.46 – 0.74]	0.99 [0.93 – 1.00]	0.63 [0.55 – 0.72]	0.97 [0.88 – 0.99]	0.68 [0.63 – 0.86]		0.80 [0.58 – 0.94]	
$S3_{hr}$	0.68 [0.51 – 0.80]	0.98 [0.93 – 1.00]	0.70 [0.61 – 0.80]	0.97 [0.88 – 0.98]	0.86 [0.67 – 0.90]		0.75 [0.59 – 0.86]	
$S3_{eda}$						0.94 [0.79 – 0.99]		0.97 [0.87 – 1.00]
$S4_{hr}$	0.61 [0.45 – 0.76]	0.98 [0.91 – 1.0]	0.59 [0.23 – 0.68]	0.93 [0.76 – 0.98]	0.72 [0.59 – 0.84]		0.84 [0.54 – 0.94]	
$S4_{eda}$						0.93 [0.82 – 1.00]		0.98 [0.91 – 1.00]
				SVM	RF			

trained on the *training dataset* was tested with an *evaluation dataset*. All SVM models were trained using the same hyper-parameters we used for LOSO cross-validation ($C = 107, \gamma = 0.001$). We found that the change in AUROC by using individually tuned parameters was negligible (less than .02 in all cases). Thus, for brevity, we only present results using constant hyper-parameters.

For comparison, along with the cross-study evaluations, we also report the LOSO cross-validation results when the *training dataset* and *evaluation dataset* were from the same study. For example, cell($S1_{math}$, $S1_{all}$) shows the median AUROC of LOSO cross-validation, where at each iteration of the cross-validation, the model was evaluated on *baseline* and *mental arithmetic*, *startle response*, and *cold pressor* tasks of the *left-out* user.

We make several observations from Table 4:

- Models built using $S1_{math}$ and $S2_{math}$ (the *baseline* and *mental arithmetic stressor* subsets of $S1$ and $S2$ respectively), performed quite well when tested with each other and resulted in similar AUROC scores. When tested using the heart-rate and R-R interval features of datasets $S3$ ($S3_{hr}$) and $S4$ ($S4_{hr}$), the AUROC scores were lower when compared with $S1$ and $S2$. Even though the participants experienced a similar stressor (mental arithmetic) in the training and test sets, we believe the discrepancy is because of the quality differences between sensors in Polar H7/H10 and Empatica E4.
- The converse, however, is more interesting. Models built using $S3_{hr}$ and $S4_{hr}$ performed quite well when tested with $S1_{math}$ and $S2_{math}$, with AUROC scores significantly higher than the models' own cross-validation scores. This suggests that even though there may be noise in the Empatica E4 data, the model was accurately able to learn the decision boundaries between *stressed* and *non-stressed* periods, thus leading to high AUROC scores when tested with $S1_{math}$ and $S2_{math}$. Further, we observe that for LOSO cross-validation of $S3_{hr}$ and $S4_{hr}$, SVM performed better than RF. However, when we applied a model built on $S3_{hr}$ to $S4_{hr}$ and vice-versa, we observe that RFs led to much better performance than SVMs.
  Further, when we included the EDA features, the performance of model trained using $S3$ on $S4$ (and vice-versa), improved significantly. The performance with EDA features is similar to the cross-validation

performance of the individual models. As in Section 5.1, we observe that RFs performed significantly better than SVM when the EDA features were included along with the heart-rate and R-R interval features.

- We observe that all models performed poorly when tested with $S1_{all}$ and $S2_{all}$. One potential reason could be that all models were tested only using the mental-arithmetic task as the stressor, whereas $S1_{all}$ and $S2_{all}$ subsets include two additional stressors: startle response and cold pressor, and models built using just mental arithmetic task are not able to identify the physiological response due to the other tasks. Further, Mishra et al. reported that in their study ($S1$), participants' perception of stress was different for the different tasks [47]. Most participants did not find the startle response test and the cold-pressor test to be as stressful as the mental-arithmetic task. This observation might help explain our results. It is possible that during the mental-arithmetic task, participants experience a physiological response which is significantly higher than the physiological response at the baseline rest condition. During the other stressors, the participants' physiological response may not be as profound as the response to the mental arithmetic task. This could mean that models built using the mental-arithmetic task and baseline rest have distinct boundaries, which in turn results in high performance of all models on $S1_{math}$ and $S2_{math}$.

Next, we discuss methods for improving the efficacy of the stress-detection models discussed in this section.

5.3 Improving Model Performance

Until now, all research in the domain of physiological stress sensing has focused on calculating some physiological features (HRV, ECG, EDA, RIP, etc.) in a given time window and building models to detect "stress" in that window. To simplify model building and evaluation, most prior works assume that these time-windows are independent, thus missing out on the temporal characteristics of stress. Stress (or the physiological response to stress) does not randomly change with each time-window; instead, it exhibits some temporal characteristics: gradual increase, gradual decrease, staying constant. Rarely are there rapid fluctuations across consecutive time windows. In this section we evaluate whether the temporal characteristics of physiological signals can be used to improve stress-detection models.

Before we discuss the different methods to account for temporal characteristics, however, it is important to address the different circumstances where stress-detection models might be useful, because our approach could potentially change with type of application. We broadly categorize the application of stress-detection models into three groups:

- In-the-moment detection: This is probably the most discussed *potential* application of stress detection models: to enable real-time in-the-moment detection. Accurate in-the-moment detection of stress could enable delivery of Just-in-Time Adaptive Interventions (JITAI) to help people manage their stress levels [59, 61].
- Batch detection of past stress: This is a less-frequently discussed application of stress detection, which we argue is also important. In this scenario, a clinician or therapist might be interested in a quantitative understanding of the stress episodes experienced by a person or a patient with anxiety disorder over a period of time; e.g., for the past month, what was the frequency, duration, and recovery period for each stressful episode? This could help clinicians customize treatment options for that patient.
- Stress prediction: This is an ambitious application in which the goal is to predict *future* stress episodes to enable delivery of an intervention even before the person experiences stress. However, given that physiological signals are non-stationary, it may not be feasible to forecast precise stress periods based on physiological signals alone. In the past, Umematsu et al. used a combination of smartphone sensors, surveys, and physiological signals (EDA and skin temperature) to forecast the "overall" stress in the evening of the next day [67].

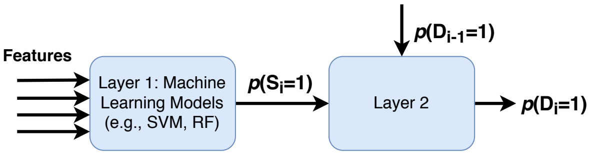

Fig. 4. Visualization of a two-layer approach for stress detection in a given time window *i*.

In our work, we discuss methods to improve stress-detection models focused on the first two scenarios: in-the-moment detection and batch detection of past stress. We hope to leverage different temporal characteristics for the two application scenarios. To detect stress in-the-moment, we can leverage information about prior stress to inform in-the-moment stress. For the second application, where the goal is to detect stress in a previously collected data, we can leverage the stress classifications *around* the current window (before and after) to better inform the stress level in the current window.

5.4 In-the-moment Detection

We discuss and evaluate two different approaches towards incorporating information about *prior* stress: a time-series approach and a stochastic-process approach (initially proposed by Mishra et al. [47]). In either case, the eventual outcome is a two-layer approach that for each time window decides on a final *detected* stress state by fusing the *sensed* stress state (for the current time window) from a regular machine-learning model, along with the *detected* stress state of the previous time window. Figure 4 illustrates this two-layered approach.

To leverage the temporal characteristics of the stress signal, we had to slightly modify our evaluation process. In Section 4.5, we discuss how we extracted windows of time from a physiological signal, labeled them as '1' (*stressed*) or '0' (*not-stressed*) to train a machine-learning model, which was then used to classify similarly extracted windows of physiological signals from other users (LOSO or cross-study). While the model-building part of the process stays the same, we no longer extracted windows of physiological signals from the test users. To use the temporal characteristics of stress, we classified the entire time-series signal of the test users in 60-second windows, rather than using only those time windows where the participant was in a stressful or rest condition. The output of the model was then a time series of *stress probability* numbers. For evaluation, we compared the performance of the models in the windows of time for which we have a ground-truth label, i.e., windows when the participants were experiencing a stressor or undergoing a rest period.

5.4.1 Time-series Approach. We used exponential smoothing [24] to incorporate the effect of prior stress to detect current stress. Although exponential smoothing is typically used for time-series forecasting, we can also leverage it for our use case. As shown in Figure 4, let us assume that for a time window i, the s_i is the probability of stress from the machine-learning model, and d_i is the final stress probability from the second layer; then, using exponential smoothing, we can define a simple recurrence relationship,

$$d_i = \alpha s_i + (1 - \alpha) d_{i-1} \tag{1}$$

where d_{i-1} is the final detected stress probability from the previous window, and $0 < \alpha < 1$. To initialize the recurrence, we set $d_0 = s_0$. We used a linear search to optimize for α based on LOSO cross-validation performance.

5.4.2 Stochastic-process Approach. We used a Bayesian network model to account for the detected stress state from the previous window along with the sensed stress state (from layer 1 model) in the current time window to output the final detected stress state for the current window. We used the approach proposed by Mishra et al., which had significant performance improvements over a single-model approach in their experiments.

Table 5. Conditional Probability Table (CPT) for the Bayesian Network model

		D_i	
D_{i-1}	S_i	0	1
0	0	1	0
0	1	$1-\gamma$	γ
1	0	$1-\delta$	δ
1	1	0	1

For a given window i, the Bayesian network model formalizes a recursive relationship between the sensed stress state in the current window (S_i), and detected stress state in the previous window (D_{i-1}) to detect the final stress state in (D_i), such that for any given window, we need to calculate $p(D_i|D_{i-1}, S_i)$.

To simplify the parameterization of $p(D_i|D_{i-1}, S_i)$, Mishra et al. made the following assumptions: (1) if the binary stress state at S_i and D_{i-1} are both *true*, then set the final detected stress (D_i) as *true*; and (2) if the binary stress states at S_i and D_{i-1} are both *false*, then set D_i as *false* [47]. These assumptions are logical and help simplify the model to two parameters, γ and δ, as shown in the conditional probability table in Table 5.

The joint probability distribution for our model is

$$P(D_i, D_{i-1}, S_i) = P(D_i|D_{i-1}, S_i) \cdot P(D_{i-1}) \cdot P(S_i) \tag{2}$$

Next, we can marginalize the $p(D_i = 1)$ at any given window i from the joint distribution as

$$p(D_i = 1) = \sum_{k,l=\{0,1\}} p(D_i = 1|D_{i-1} = k, S_i = l) \cdot p(D_{i-1} = k) \cdot p(S_i = l) \tag{3}$$

Considering $p(D_i = 1)$ as y_i, and $p(S_i = 1)$ as x_i, along with the CPT in Table 5, we can simplify the above equation as the following recurrence relation

$$p(D_i = 1) = y_i = \gamma(1 - y_{i-1})x_i + \delta y_{i-1}(1 - x_i) + y_{i-1}x_i \tag{4}$$

At $i = 0$, we assume D_0 is the same as S_0, thus the recurrence can be initialized as

$$p(D_0 = 1) = y_0 = p(S_0 = 1) = x_0 \tag{5}$$

Now, using Equations 4 and 5, $p(D_i = 1)$ can be marginalized for any window i. Here γ and δ can be treated as hyper-parameters, which we tuned to maximize LOSO cross-validation performance. We used the Hyperopt [4] package to optimize parameters.

5.4.3 Model Comparison. We can now compare the efficacy of both approaches using LOSO cross-validation. We built models for each of the eight data subsets, like in Section 5.1. Initially we did a parameter search for each method – α for exponential smoothing, and γ and δ for the Bayesian network model – on each data subset. In both cases, however, we noticed that the optimal parameters were quite similar across all the data subsets. The linear search revealed that α fell between 0.54 and 0.55 for all eight subsets, so we chose $\alpha = 0.54$ for reporting results from all the models based on exponential filtering. The parameter optimization for γ and δ revealed that for all models, $\gamma \in [0.31, 0.34]$ and $\delta \in [0.85, 0.87]$. Again, to simplify model building, comparison, and reporting, we set $\gamma = 0.33$ and $\delta = 0.86$ for all data subsets. We found only minor differences in performance between our chosen hyper-parameters and the optimal parameters for each dataset.

We list the median AUROC scores for LOSO cross-validation using both the exponential filtering approach and the Bayesian network model in Table 6. These results can be compared with the LOSO cross-validation results reported in Table 4, since the layer 1 models are exactly the same ones used for the respective cross-validation in Section 5.2. For each evaluation, we also list the absolute change in median AUROC from Table 4. It is evident from Table 6 that both two-layer approaches led to a performance increase over a single-layer approach. Overall, the Bayesian network model leads to a higher improvement in performance than the exponential filtering model.

Next, we build the two-layered Bayesian network models for each subset and evaluate with another subset for cross-study evaluation. We summarize the results in Table 7. The results presented in each cell of Table 7 can directly be compared with the results in Table 4 to evaluate the efficacy of the two-layer approach with Bayesian

Table 6. LOSO cross-validation results for the different data subsets using the two-layered stress detection approach. The cells represent the median AUROC from LOSO cross-validation for each data subset. The round brackets highlight the absolute increase in AUROC score when compared with the LOSO cross-validation results in Table 4; the square brackets show the IQR.

	$S1_{math}$	$S1_{all}$	$S2_{math}$	$S2_{all}$	$S3_{hr}$	$S3_{eda}$	$S4_{hr}$	$S4_{eda}$
Exponential filtering	1.00 (+0.02) [0.98 − 1.00]	0.70 (+0.03) [0.54 − 0.93]	0.97 (+0.00) [0.86 − 1.00]	0.64 (+0.01) [0.53 − 0.71]	0.89 (+0.03) [0.66 − 0.97]	0.98 (+0.04) [0.81 − 1.00]	0.87 (+0.02) [0.55 − 0.98]	1.00 (+0.02) [0.95 − 1.00]
Bayesian network model	1.00 (+0.02) [0.99 − 1.00]	0.83 (+0.16) [0.61 − 1.00]	0.98 (+0.01) [0.91 − 1.00]	0.76 (+0.13) [0.62 − 0.81]	0.91 (+0.05) [0.70 − 0.98]	0.96 (+0.02) [0.86 − 0.99]	0.95 (+0.11) [0.73 − 0.98]	1.00 (+0.02) [0.92 − 1.00]

Table 7. Cross-study evaluation using the two-layered approach with Bayesian network model: Each cell represents the median AUROC for a model built on the training set and tested on the evaluation set. The round brackets highlight absolute increase in AUROC score when compared with the same cells in Table 4. The Interquartile Range (IQR) is shown in square brackets. In case the training and evaluation datasets are from the same study, we report the results of a LOSO cross-validation. The blue cells show results from a SVM model; the yellow cells show results from a Random Forest (RF) model. As data from studies $S1$ and $S2$ did not include any EDA data, models built with the combination of heart-rate, R-R interval, and EDA features – specifically $S3_{eda}$ and $S4_{eda}$ – could only be evaluated and compared with models from the $S3$ and $S4$ studies.

Training Dataset	***Evaluation Dataset***							
	$S1_{all}$	$S1_{math}$	$S2_{all}$	$S2_{math}$	$S3_{hr}$	$S3_{eda}$	$S4_{hr}$	$S4_{eda}$
$S1_{math}$	0.83 (+0.16) [0.54 − 0.93]	1.00 (+0.02) [0.98 − 1.00]	0.81 (+0.19) [0.65 − 0.95]	0.98 (+0.05) [0.93 − 1.00]	0.87 (+0.12) [0.56 − 0.94]		0.84 (+0.04) [0.60 − 0.97]	
$S2_{math}$	0.73 (+0.12) [0.55 − 0.98]	1.00 (+0.01) [0.96 − 1.00]	0.76 (+0.14) [0.62 − 0.81]	0.98 (+0.01) [0.91 − 1.00]	0.80 (+0.12) [0.55 − 0.92]		0.84 (+0.04) [0.63 − 0.96]	
$S3_{hr}$	0.85 (+0.12) [0.75 − 0.99]	0.99 (+0.01) [0.95 − 1.00]	0.84 (+0.14)) [0.73 − 0.98]	0.99 (+0.02)) [0.92 − 1.00]	0.91 (+0.05) [0.70 − 0.98]		0.89 (+0.14) [0.62 − 0.96]	
$S3_{eda}$						0.96 (+0.02) [0.86 − 0.99]		0.99 (+0.02) [0.88 − 1.00]
$S4_{hr}$	0.80 (+0.19) [0.63 − 0.94]	1.00 (+0.02) [0.96 − 1.00]	0.74 (+0.15) [0.67 − 0.85]	0.99 (+0.06) [0.96 − 1.00]	0.80 (+0.08) [0.55 − 0.93]		0.95 (+0.11) [0.73 − 0.98]	
$S4_{eda}$						0.95 (+0.02) [0.88 − 1.00]		1.00 (+0.02) [0.92 − 1.00]
				SVM	RF			

network models. The machine-learning models (SVM or RF) in layer 1 remained unchanged between the two evaluations. From Table 7, we observe that the two-layered approach improved the AUROC score for all models. Thus, the improvement is observed not just for LOSO cross-validation of an individual model, but also when a model is tested with data from different studies.

A drastic improvement is observed for the models that performed poorly in Table 4, suggesting that the two-layered approach improved the base classifier's ability to separate between the *stressed* and *not stressed* classes. For instance, when we trained an SVM using $S1_{math}$ and tried to detect stressful episodes in $S2_{all}$ – which included the mental arithmetic, startle response, and cold-pressor stressors – the AUROC score of the model was 0.62. Using the two-layer approach, however, the performance improved to 0.81 – an improvement of over 31%. On average, the two-layered approach led to an AUROC improvement of 0.07 over a single-model approach.

5.5 Batch Detection of Past Stress

In the previous section, we evaluated the potential of leveraging prior stress to detect *in-the-moment* stress.

Our results suggest that performance improvement due to a two-layered approach is consistent across studies, irrespective of the device type, sensor type, or the type of stressors. There could, however, be applications where the goal is to detect stress instances or episodes on some previously collected data. For such cases, we explore the feasibility to leverage stress classifications *around* the current window, i.e., before and after the current window to better inform the stress level in the current window. The assume that since the data has already been collected, we can iterate over the signals to make a robust classification.

Since a stochastic-process approach worked better than a timeseries approach for *in-the-moment* detection, we decided to try a stochastic-process approach here as well. To this end, we expanded the Bayesian network model discussed in Section 5.4.2. We now include another stress state, S_{i+1}, which is the sensed stress state at time $i+1$. Hence, now for any given window, we need to calculate $p(D_i | D_{i-1}, S_i, S_{i+1})$. Again, we make some assumptions to simplify the model to four parameters: α, β, γ, and δ, as illustrated in the CPT in Table 8.

Table 8. Condition Probability Table (CPT) for the extended Bayesian network model

			D_i	
D_{i-1}	S_i	S_{i+1}	**0**	**1**
0	0	0	1	0
0	0	1	$1-\alpha$	α
0	1	0	1	0
0	1	1	$1-\beta$	β
1	0	0	$1-\gamma$	γ
1	0	1	0	1
1	1	0	$1-\delta$	δ
1	1	1	0	1

Following an approach similar to Section 5.4.2, we were able to marginalize $p(D_i = 1)$ from the joint distribution $P(D_i, D_{i-1}, S_i, S_{i+1})$, and simplify it as

$$p(D_i = 1) = y_i = x_{i+1}(1 - y_{i-1})[\alpha(1 - x_i) + \beta x_i] + y_{i-1}[(1 - x_i)(\gamma(1 - x_{i+1}) + x_{i+1}) + x_i(\delta(1 - x_{i+1}) + x_{i+1})] \quad (6)$$

where $y_i = p(D_i = 1)$ and $x_i = p(S_i = 1)$. Further, for a signal with n time windows, we assume $y_0 = x_0$ and $y_n = x_n$.

We used the Hyperopt [4] package to tune the four parameters by optimizing for AUROC in LOSO cross-validation. We found that, unlike the approach discussed in Section 5.4.3, we could not find a common range for the hyperparameters α, β, γ, and δ that performed well for each cross-validation. In fact, our results showed that while γ and δ were consistent across each model, the α and β values varied significantly. We report the results obtained by a LOSO cross-validation for each data subset in Table 9 using the optimal hyperparameters for each model.

We found that – even with optimal parameters – the performance was marginally better than the performance achieved in Section 5.4.2. It seems that the likelihood of stress in the current window is more influenced by the likelihood of stress in the previous window, as compared to the next window. Thus, given the lack of substantially better results with this approach, along with the added complexity of having different hyperparameter for each model, we did not conduct a cross-study evaluation for this approach.

Table 9. LOSO cross-validation results for the different data subsets for "batch detection of past stress" using Bayesian network model. The cells represent the median AUROC from LOSO cross-validation for each data subset. The round brackets highlight the absolute difference in AUROC score when compared with the LOSO cross-validation results of the Bayesian network model in Table 6; the square brackets show the IQR.

	$S1_{math}$	$S1_{all}$	$S2_{math}$	$S2_{all}$	$S3_{hr}$	$S3_{eda}$	$S4_{hr}$	$S4_{eda}$
Batched Bayesian	1.00 (+0.00)	0.85 (+0.02)	0.99 (+0.01)	0.74 (-0.02)	0.92 (+0.01)	0.98 (+0.02)	0.96 (+0.01)	1.00 (+0.00)
network model	[0.97 – 1.00]	[0.61 – 1.00]	[0.93 – 1.00]	[0.63 – 0.77]	[0.68 – 0.98]	[0.84 – 0.99]	[0.69 – 0.99]	[0.93 – 1.00]

5.6 Finding Threshold for Classification

Having improved the performance of the machine-learning models, we now discuss the actual decision-making component, i.e., when should the model classify an instance as *stressed*? Either a binary (*not stressed*, *stressed*) or a ternary (*low*, *medium*, *high*) output are over-simplifications of stress detection, a practice consistent across prior works that empirically evaluate stress-detection models [28, 38, 39, 47, 52, 61, 66]. Most machine-learning classifiers used for detecting stress, including SVM and RF, output a probability of stress. By default, the classifiers assume a 0.5 threshold, such that output instances with probability greater than 0.5 are classified as *stress*. In the past, researchers have used a custom threshold to determine when an instance is classified as *stress* [38, 47, 61]. To find such a threshold, the usual approach has been to tune it like a hyper-parameters using LOSO cross-validation that optimizes some performance metric, usually the F1-score [38, 47]. This method works relatively well in homogeneous datasets in which each participant's data has a similar distribution because they all underwent the same study procedure. As shown in Table 3, however, the optimal threshold can vary across datasets.

To dig deeper, we evaluated the models built for each dataset ($S1_{math}$, $S2_{math}$, $S3_{hr}$, $S4_{hr}$) on the $S2_{all}$ dataset. For each model, we searched for the optimal threshold that maximized the F1-score. We report the results in Figure 5. The immediate observation from Figure 5 is the improved performance, which is even better than the LOSO performance of just a single model reported in Table 3, once again highlighting the efficacy of the two-layered approach. The more important observation, however, is the drastic variation in thresholds. Even models built on $S1_{math}$ and $S2_{math}$, which had similar sensor types and study protocol, there was a variation in the optimal threshold. To gain more insight, we explored the *minute-by-minute* stress likelihood scores of each model on a random participant X from $S2$, which we present in Figure 6. We observe that this participant started with some residual stress that declined during the baseline period. During the startle response test (T1), we made a loud sudden noise behind the participant, which may have created the first spike in the likelihood of stress; we made another loud noise just before the end of T1, which may have caused the spike in stress likelihood at the start of the recovery period. Next, during the mental arithmetic task (T2), the likelihood of stress stayed elevated for the entirety of the session, and started to decline during the recovery period. Finally, for the cold pressor test (T3) the participant's stress levels spiked initially, but (as they became accustomed to the cold) the stress level started to recover, with it finally returning to baseline during the final recovery period. Further, we observe that the trends for the models seem similar, albeit with different amplitudes. The baseline stress likelihood from S_{hr} seems to be more elevated than the others. Although there is a separation between the baseline and stress periods, it is not as pronounced as in the other models. This higher baseline is why the classification threshold for $S4_{hr}$ was significantly higher than the other models; although the other models are similar they do vary slightly in their amplitudes, resulting in different "optimal" thresholds.

The fact that different models can have different optimal thresholds – and that the threshold can depend on the test dataset – makes it challenging to deploy or even test models built using one dataset on independent and unseen *test* users with no a priori knowledge of the data distribution. For real-world use, the threshold cannot be tuned for each dataset or user since no ground-truth labels will be available. Hence, we believe the only possible solutions are to use a "fixed" threshold selected using the training data – and sacrifice performance – or to develop an adaptive or unsupervised method to select the threshold for new users or groups.

We next explore one such unsupervised method. The AUROC score for most of the models we evaluated was high, which suggests that there exist a high level of separation between the *stressed* and baseline rest conditions. We thus explore the feasibility of clustering methods to assign *stressed* and *non-stressed* labels to our data. To this end, we configured a K-Means algorithm with two clusters; we initialized the clusters' centers at 0 and 1, to denote the *rest* and *stressed* periods, respectively. We then fit the K-Means algorithm with the stress probabilities (predicted by a stress-detection model) of each participant. Once the clustering was complete, we did a final iteration through each minute (four stress probabilities) and if the four probabilities belonged to different clusters, we assigned all four points to the majority cluster; in case of a tie, we assigned all four points to the cluster of the previous minute. We fit a different K-Means model to each participant; hence, the final cluster centers could be different for each user.

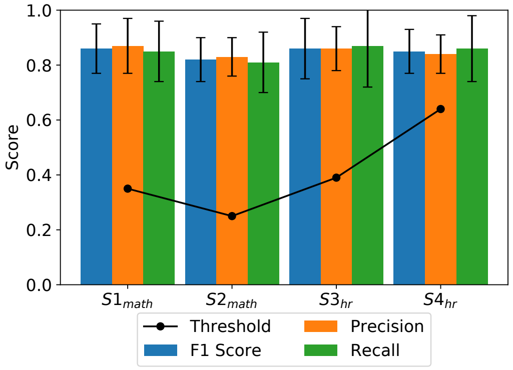

Fig. 5. Performance comparison for the different models on $S2_{all}$, along with the respective classification thresholds.

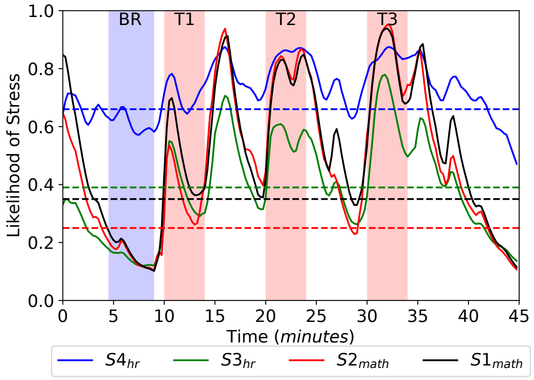

Fig. 6. Stress likelihood estimated by the different models on a participant from $S2$. The region marked as BR represents the baseline rest period, region $T1$ represents the *startle response* task, region $T2$ represents the *mental arithmetic* task, and $T3$ represents the *cold pressor* task. The dashed-lines represent the model specific thresholds (same as Figure 5).

To evaluate the efficacy of a cluster-based classification, we compare the results of the clustering with two thresholds: a 'fixed threshold' learned during the model's training phase, and an 'optimized threshold' learned from the testing data. In real-world stress detection applications, however, an 'optimized' threshold cannot be learned as the test data has no ground-truth labels. We report the F1-score, precision, and recall metrics for each model in Tables 10 & 11. In situations where the training dataset and evaluation dataset were exactly the same, the optimized threshold was the same as the fixed threshold.

We found the performance of the clustering-based approach always resulted in better classification performance than the 'fixed threshold', and at times even better than the 'optimized threshold'. Since choosing an optimal threshold using each *test* set violates the independence of data, and since no labels will be available when stress-detection models are deployed, we recommend this clustering-based classification approach be used for deployments of stress-detection models. Further, we observe that under certain situations, e.g., training on $S1_{math}$ and testing on $S2_{all}$, the clustering-based approach resulted in high precision, but relatively low recall, suggesting that the model was not able to identify all the stressful periods in $S2$, but the ones it did identify were very accurate.

To visualize how the clusters are assigned to the signals, we plot in Figure 7 the final results (after clustering) of applying $S1_{math}$ and $S4_{hr}$ models to the same $S2$ participant X we saw in Figure 5. The plots provide help explain why the recall from the clustering-based approach may have been low. In the case of this participant, it seems that the model was not able to identify all the stress-minutes during the startle response task (T1), whereas an optimized threshold of 0.35 was able to classify all of the minutes of T1 as stressed.[7] Given the trajectory of the *stress signal* – in the case, of user X – it seems that the clustering approach might be more informative, as it is able to distinguish periods when the users' stress levels are going down, and marks them as such. This

[7] For clarity, we show the output for just two models. We argue, however, that since $S1_{math}$ had the most variance (from Figure 6) and $S4_{hr}$ had the least variance, these two models should provide a good representation of the effectiveness of our approach.

Table 10. Performance of the models built using heart rate and R-R interval features based on the type of threshold. 'FT' denotes a fixed threshold learned during training on the train set; 'OT' denotes an optimized threshold learned from the test set; 'US' denotes the proposed unsupervised clustering based classification. $F1$, P, and R represent the F1-score, precision, and recall, respectively.

Training Dataset		***Evaluation Dataset***																	
		$S1_{math}$			$S1_{all}$			$S2_{math}$			$S2_{all}$			$S3_{hr}$			$S4_{hr}$		
		FT	OT	US	FT	OT	US	FT	OT	US	FT	OT	US	FT	OT	US	FT	OT	US
$S1_{math}$	$F1$	0.89		0.89	0.68	0.85	0.71	0.84	0.86	0.88	0.70	0.86	0.76	0.65	0.67	0.68	0.72	0.75	0.80
	P	0.88		0.89	0.93	0.83	0.92	0.85	0.85	0.84	0.92	0.87	0.91	0.62	0.60	0.62	0.74	0.74	0.77
	R	0.89		0.89	0.53	0.87	0.58	0.83	0.88	0.92	0.57	0.85	0.65	0.69	0.76	0.76	0.71	0.77	0.83
$S2_{math}$	$F1$	0.85	0.87	0.86	0.66	0.84	0.70	0.87		0.89	0.68	0.82	0.72	0.73	0.74	0.75	0.75	0.76	0.79
	P	0.82	0.85	0.84	0.89	0.83	0.88	0.86		0.85	0.90	0.83	0.90	0.62	0.62	0.65	0.68	0.68	0.72
	R	0.88	0.88	0.88	0.52	0.85	0.58	0.88		0.94	0.54	0.81	0.60	0.89	0.92	0.88	0.83	0.86	0.87
$S3_{hr}$	$F1$	0.82	0.90	0.91	0.57	0.90	0.70	0.81	0.93	0.92	0.58	0.86	0.74	0.77		0.76	0.73	0.74	0.80
	P	0.93	0.89	0.96	0.93	0.90	0.97	0.97	0.93	0.92	0.98	0.86	0.95	0.70		0.68	0.64	0.64	0.72
	R	0.73	0.90	0.86	0.41	0.90	0.55	0.70	0.94	0.93	0.41	0.87	0.61	0.85		0.86	0.86	0.88	0.91
$S4_{hr}$	$F1$	0.88	0.89	0.86	0.75	0.84	0.79	0.86	0.87	0.86	0.75	0.85	0.82	0.69	0.74	0.74	0.77		0.89
	P	0.85	0.90	0.79	0.90	0.81	0.87	0.82	0.85	0.79	0.89	0.84	0.86	0.63	0.62	0.63	0.76		0.70
	R	0.91	0.87	0.95	0.64	0.87	0.73	0.92	0.90	0.95	0.65	0.86	0.78	0.76	0.91	0.89	0.78		0.91

Table 11. Performance of the models built using EDA features based on the type of threshold. 'FT' denotes a fixed threshold learned during training on the train set; 'OT' denotes an optimized threshold learned from the test set; 'US' denotes the proposed unsupervised clustering based classification. $F1$, P, and R represent the F1-score, precision, and recall, respectively.

Training Dataset		***Evaluation Dataset***					
		$S3_{eda}$			$S4_{eda}$		
		FT	OT	US	FT	OT	US
$S3_{eda}$	$F1$	0.80		0.81	0.85	0.86	0.88
	P	0.79		0.78	0.89	0.86	0.92
	R	0.82		0.84	0.81	0.85	0.84
$S4_{eda}$	$F1$	0.81	0.81	0.81	0.89		0.84
	P	0.75	0.77	0.76	0.89		0.87
	R	0.89	0.85	0.87	0.88		0.82

example emphasizes why it is important to avoid relying only on metrics like F1-score, precision, and recall, which provide limited information about a model's performance.

Further, we believe that focusing only on two classes (*stressed* and *not-stressed*) is too much of an over-simplification. We argue that there should at least be three classes: *not-stressed*, *intermediate*, and *stressed*. In the future, such a three-class classification could enable scheduling of Just-in-time Adaptive Interventions (JITAI). In one potential use case, a JITAI system would trigger interventions when the probability of stress is in the *intermediate* stage with a positive slope.

We next explore the feasibility of such a three-class scenario. We now fit a K-Means model with 3 clusters, with the centers initialized at 0, 0.5, and 1.0 for the *baseline*, *intermediate*, and *stressed* classes, respectively. We used an approach to fit and interate over the cluster much like that we used for the two-class approach. Given the preliminary scope of this three-class exploration, we quickly revisit participant X in Figure 8. The three-class approach appears promising, and worthy of further investigation.

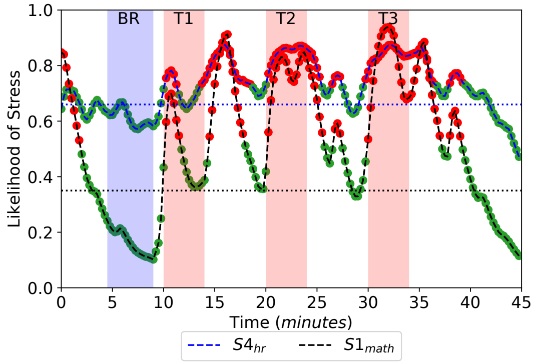

Fig. 7. The two-class cluster assignments for two different models, $S4_{hr}$ and $S1_{math}$, on a participant from $S2$. The green dots represent the "not-stressed" cluster, and the red dots represent the "stressed" cluster. The dotted-lines represent the optimal thresholds for the two models (same as Figure 5).

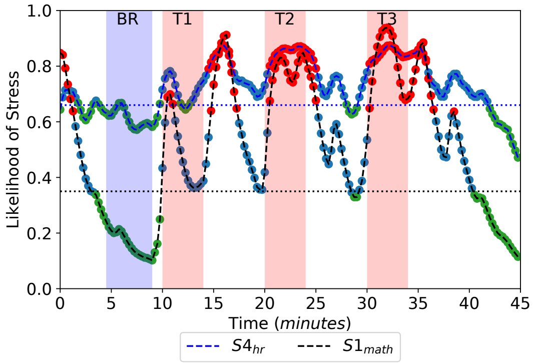

Fig. 8. The three-class cluster assignments for two different models, $S4_{hr}$ and $S1_{math}$, on a participant from $S2$. The green dots represent the "baseline" cluster, the blue dots represent the "intermediate" cluster, and the red dots represent the "stressed" cluster. The dotted-lines represent the optimal thresholds for the two models (same as Figure 5).

It is important to note that the clustering-based approach works only for batch detection of past stress, and not for detecting stress *in-the-moment*. We argue, however, that this method can be used to learn the cluster-centers of the user for a few days before enabling an in-the-moment stress-detection module based on the learned cluster centers. In any case, there remain other challenges before a real-time *in-the-moment* stress-detection method is feasible, as we discuss in Section 6.

5.7 Summary of Results

We now summarize the results presented in Section 5. We started with a LOSO cross-validation of each data subset and found that even after following the same data cleaning, processing, feature extraction, and model building process, the results of each LOSO evaluation was different, even between studies that had similar study protocols or sensors. We specifically noted that models built with just the mental arithmetic task did not perform as well when tested with the subsets consisting of all three stressors in S1 and S2. We also observed that models built using only the heart-rate and R-R interval features from the Empatica E4 performed poorly as compared to models built using data from the Polar H7/H10. Finally, we noted that the optimal threshold to classify between *stressed* and *not-stressed* was different across the studies, and the values of the F1 score, precision, and recall varied significantly with a change in threshold.

Next, we built models using data from each study, and evaluated it on every other subset. We found that models built using $S1_{math}$ and $S2_{math}$ (the *baseline* and *mental arithmetic stressor* subsets of $S1$ and $S2$) performed similarly when tested with the other's data. Both these studies had similar protocols (except the random order of stressors) and used the same family of Polar heart-rate monitors. These models, however, had lower AUROC scores when evaluated with data from $S3$ and $S4$, although the type of stressor was similar (mental arithmetic). We believe this difference may result from signal-quality differences between Polar and Empatica devices. The models built using heart-rate and R-R interval data from $S3$ and $S4$, however, performed quite well when tested with $S1$ and $S2$. This result suggests that – despite the lower data quality from Empatica – these SVM models were able to reliably fit a hyperplane between the features belonging to *stressed* and *not-stressed* periods. Finally, we observed that all models performed poorly when tested with $S1_{all}$ and $S2_{all}$. One possible explanation: our models were built using just the baseline rest period and the mental arithmetic stressor, but $S1_{all}$ and $S2_{all}$ included data from other stressors that may have triggered a different physiological response.

We then explored means to improve the performance of these stress-detection models by accounting for the temporal dynamics of stress signals. We found that a two-layer approach with a Bayesian-network model could factor in the previous window's stress levels, resulting in a consistent improvement in performance – across all models – when compared to a single-layer approach. We observed an average increase in AUROC of 0.07. The increase was more pronounced for evaluations that performed poorly before; e.g., applying the two-layer approach to evaluate model built with $S1_{math}$ on $S2_{all}$ led to an increase of 0.19 in the median AUROC.

We then explored means to determine the threshold for classification. We found that the choice of threshold varied with the choice of sensors, study protocol, and distribution of train and test samples. We proposed an unsupervised clustering approach to determine the *stressed* and *not-stressed* classes , and found that our method always performed better than when using a threshold learned from training data. Finally, we presented some qualitative evidence showing the feasibility of clustering the data into three classes: *baseline*, *intermediate*, and *stress*.

6 DISCUSSION AND FUTURE DIRECTIONS

In this paper we evaluate the reproducibility of stress-detection models. We identified several challenges that might limit reproducibility, and discussed and proposed methods to tackle those challenges: improving the performance of models across studies and determining a threshold for classification that could be applicable across studies. While we argue that these are important steps in the direction of generalizability of stress detection models, several challenges and limitations remain; we discuss them here.

6.1 Detecting Stress "in-the-moment"

As part of our work, we evaluated methods to improve detection of stress *in-the-moment*; while the results were positive, there are some considerations. First, we argue it is not yet feasible to deploy real-time, in-the-moment stress detection that works immediately out-of-the-box. One of the key processing steps in all stress-detection approaches is the *normalization* of the physiological signals for each user, to remove any user-specific traits from the signal. To normalize a participant's signal, we (and prior work) used the entire time series of data from that participant. So although the stress classification was happening *in-the-moment*, the signal processing and feature computations were done beforehand. For a stress-detection method to work out-of-the-box, the model needs to do real-time data normalization of physiological signals before features can be calculated. Since the mean and standard deviation of such signals could vary over a day (or a few days), these models would have to employ a form of adaptive normalization to make reliable predictions. Adaptive normalization, however, is a non-trivial task and in itself is an important topic of research [49, 50].

For the deployment of just-in-time stress detection, the models will likely need a calibration/personalization period lasting several days. During this adaptation period, the model could learn the normalization parameters and classification thresholds for the new user (based on our proposed clustering-based classification method), without requiring any labels or feedback from the user.

6.2 Detecting Stress in "Free-Living" Conditions

We recognize that our work focuses on data collected in controlled in-lab settings. Given the lack of reproducibility and replicability of results in a controlled scenarios, however, we argue it is important to evaluate reproducibility in a controlled environment before tackling reproducibility in free-living conditions. Our work provides a solid foundation for such work.

Further, we argue that physiological sensors alone may not be sufficient for detecting stress in free-living conditions. These sensors simply measure the physiological response to stress. A variety of other factors could cause similar physiological responses, e.g., caffeine intake, smoking, drinking, and physical activity, all of which can confound stress-detection models. In fact, Grace et al. found that listening to music (which had a relaxing effect on participants' perception of stress) led a physiological arousal that could be misconstrued for stress [10]. Thus, we argue that for continuous real-time stress detection to be feasible in free-living conditions, physiological signals need to be paired with the users' *context*. Although some prior works have accounted for physical activity in their stress-detection models with varying degrees of success [28, 38, 47, 61], few studies have addressed the other confounding factors. In a preliminary study, Mishra et al. found that the threshold of what is considered 'stress' changed with participants' context [45]. The authors reported that during inherently "arousing" activities, like attending a meeting or driving, the threshold of physiological stress likelihood for participants to perceive a situation as stressful was higher than restful activities. Given these and other remaining challenges, free-living stress detection remains an active area of research.

6.3 Large-scale Validations

In our work, we evaluated the reproducibility of stress detection models using data from 90 participants collected from four independent studies conducted over a period of two years. The studies had different participant demographics, different study protocols, and different physiological sensors. While our results are promising – demonstrating methods and models that can be translated across studies – further evaluations with larger studies are needed. One important aspect is the accuracy of the sensing devices. We found that less-accurate sensors could lead to weaker performance, regardless of the modeling quality. Hence, when evaluating performance of models across studies involving different sensors, we recommend comparing performance with the individual study's LOSO cross-validation scores. Performance comparisons that use the same benchmark across studies with differing sensors may lead to inflated or deflated results and incorrect conclusions.

6.4 Stress is not a Binary Condition

Researchers have long treated stress detection as a binary classification problem [17, 38, 39, 47, 52, 61, 66]. Although this approach simplifies the model complexity and improves detection performance, one must remember that 'stress' is a continuous variable and is continuously varying. Few researchers have explored stress detection with more than two classes [28, 35, 48]. One approach is to build a binary classifier but use the model's output (probability of stress) as a continuous variable to measure stress response. In our work, we considered this probability as a continuous stress signal and evaluated the potential for use of an unsupervised method to identify periods of *stressed* and *not-stressed*. We also provided preliminary qualitative evidence indicating it may be feasible to group the stress signal into three categories: *baseline*, *intermediate*, *stress*. Such a multi-class approach would be vital to delivering just-in-time adaptive interventions to users. Although Sarker et al. discuss a complex trend-analysis method [61], we expect a simple heuristic approach, like delivering interventions when the stress level is in the intermediate class with a positive slope, may well be sufficient. We plan to further evaluate this method in future work.

7 CONCLUSION

In this work, we took the first step towards testing reproducibility and validity of methods and machine-learning models for stress detection. We analyzed data from 90 participants, from four independent controlled studies, using two different types of sensors, with different study protocols and research goals. We started by evaluating the performance of models built using data from one study and tested on data from other studies. Next, we evaluated new methods to improve the performance of stress-detection models and found that our methods led to a consistent increase in performance across all studies, irrespective of the device type, sensor type, or the type of stressor. Finally, we developed and evaluated a clustering approach to determine the *stressed/not-stressed* classification when applying models on data from different studies, and found that our approach performed better than selecting a threshold based on training data. This paper's thorough exploration of reproducibility in a controlled environment provides a critical foundation for deeper study of such methods, and is a prerequisite for tackling reproducibility in free-living conditions.

ACKNOWLEDGMENTS

This research results from a research program at the Center for Technology and Behavioral Health at Dartmouth College, supported by the National Science Foundation under award numbers CNS-1314281, CNS-1314342, CNS-1619970, and CNS-1619950, and by the NIH National Institute on Drug Abuse under award number NIH/NIDA P30DA029926. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the sponsors.

REFERENCES

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

[2] APA. 2019. Stress in America 2019. https://www.apa.org/news/press/releases/stress. [Online; accessed 10-Oct-2019].

[3] Brinnae Bent, Benjamin A. Goldstein, Warren A. Kibbe, and Jessilyn P. Dunn. 2020. Investigating sources of inaccuracy in wearable optical heart rate sensors. *npj Digital Medicine* 3, 1 (Dec. 2020), 1–9. https://doi.org/10.1038/s41746-020-0226-6

[4] J Bergstra, D Yamins, and D D Cox. 2013. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28 (ICML'13)*. JMLR.org, I–115–I–123.

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

[6] George Boateng, Vivian Genaro Motti, Varun Mishra, John A. Batsis, Josiah Hester, and David Kotz. 2019. Experience: Design, Development and Evaluation of a Wearable Device for mHealth Applications. In *Proceedings of the International Conference on Mobile Computing and Networking (MobiCom)*. https://doi.org/10.1145/3300061.3345432

[7] JJ Braithwaite, DG Watson, Jones Robert, and Rowe Mickey. 2013. A Guide for Analysing Electrodermal Activity (EDA) & Skin Conductance Responses (SCRs) for Psychological Experiments. *Psychophysiology* 49, 8 (2013), 1–42. https://www.birmingham.ac.uk/documents/college-les/psych/saal/guide-electrodermal-activity.pdf

[8] Gillian Butler. 1993. Definitions of stress. *Occasional paper (Royal College of General Practitioners)* 61 (1993), 1.

[9] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)* 2, 3 (2011), 1–27.

[10] Grace Chen, Varun Mishra, and Ching-Hua Chen. 2019. Temporal Factors of Listening to Music on Stress Reduction. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers (UbiComp/ISWC '19 Adjunct)*. ACM, New York, NY, USA, 907–914. https://doi.org/10.1145/3341162.3346272

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

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

[13] Luca Citi. 2020. cvxEDA. https://github.com/lciti/cvxEDA. [Online; accessed 04-February-2020; Git commit 7928444].

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

[15] Sharon M Crook, Andrew P Davison, and Hans E Plesser. 2013. Learning from the past: approaches for reproducibility in computational neuroscience. In *20 Years of Computational Neuroscience*. Springer, 73–102.

[16] Elena Di Lascio, Shkurta Gashi, and Silvia Santini. 2018. Unobtrusive Assessment of Students' Emotional Engagement during Lectures Using Electrodermal Activity Sensors. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3, Article Article 103 (Sept. 2018), 21 pages. https://doi.org/10.1145/3264913

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

[18] Emaptica. 2018. Empatica E4. https://www.empatica.com/en-eu/research/e4/. [Online; accessed 06-August-2020].

[19] Emre Ertin, Nathan Stohs, Santosh Kumar, Andrew Raij, Mustafa al'Absi, and Siddharth Shah. 2011. AutoSense: unobtrusively wearable sensor suite for inferring the onset, causality, and consequences of stress in the field. In *Proceedings of the ACM Conference on Embedded Networked Sensor Systems (SenSys)*. ACM, 274–287. https://doi.org/10.1145/2070942.2070970

[20] Association for Computing Machinery. 2016. Artifact Review and Badging. https://www.acm.org/publications/policies/artifact-review-badging. [Online; accessed 04-August-2020].

[21] H. Gao, A. Yüce, and J. Thiran. 2014. Detecting emotional stress from facial expressions for driving safety. In *2014 IEEE International Conference on Image Processing (ICIP)*. 5961–5965.

[22] Maurizio Garbarino, Matteo Lai, Simone Tognetti, Rosalind Picard, and Daniel Bender. 2014. Empatica E3 - A wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition. In *Proceedings of the 4th International Conference on Wireless Mobile Communication and Healthcare - "Transforming healthcare through innovations in mobile and wireless technologies"*. ICST. https://doi.org/10.4108/icst.mobihealth.2014.257418

[23] Enrique Garcia-Ceja, Venet Osmani, and Oscar Mayora. 2016. Automatic Stress Detection in Working Environments From Smartphones' Accelerometer Data: A First Step. *IEEE Journal of Biomedical and Health Informatics* 20, 4 (Jul. 2016), 1053–1060. https://doi.org/10.1109/JBHI.2015.2446195

[24] Everette S. Gardner. 1985. Exponential smoothing: The state of the art. *Journal of Forecasting* 4, 1 (Jan. 1985), 1–28. https://doi.org/10.1002/for.3980040103

[25] Shkurta Gashi, Elena Di Lascio, and Silvia Santini. 2019. Using Unobtrusive Wearable Sensors to Measure the Physiological Synchrony Between Presenters and Audience Members. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 1, Article Article 13 (March 2019), 19 pages. https://doi.org/10.1145/3314400

[26] Madeline Gaynor, Abbey Sawyer, Sue Jenkins, and Jamie Wood. 2019. Variable agreement between wearable heart rate monitors during exercise in cystic fibrosis. *ERJ Open Research* 5, 4 (Oct. 2019), 00006–2019. https://doi.org/10.1183/23120541.00006-2019

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

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

[29] Alberto Greco, Gaetano Valenza, Antonio Lanata, Enzo Pasquale Scilingo, and Luca Citi. 2015. cvxEDA: A convex optimization approach to electrodermal activity processing. *IEEE Transactions on Biomedical Engineering* 63, 4 (2015), 797–804.

[30] Kristina Grifantini. 2010. Sensor Detects Emotions through the Skin. https://www.technologyreview.com/s/421316/sensor-detects-emotions-through-the-skin/

[31] J. A. Hanley and B. J. McNeil. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143, 1 (1982), 29–36. https://doi.org/10.1148/radiology.143.1.7063747

[32] Tian Hao, Henry Chang, Marion Ball, Kun Lin, and Xinxin Zhu. 2018. cHRV uncovering daily stress dynamics using bio-signal from consumer wearables. In *MEDINFO 2017: Precision Healthcare Through Informatics: Proceedings of the 16th World Congress on Medical and Health Informatics*, Vol. 245. IOS Press, 98.

[33] Tian Hao, Jeffrey Rogers, Hung-Yang Chang, Marion Ball, Kimberly Walter, Si Sun, Ching-Hua Chen, and Xinxin Zhu. 2017. Towards Precision Stress Management: Design and Evaluation of a Practical Wearable Sensing System for Monitoring Everyday Stress. *iproc* 3, 1 (22 Sep 2017), e15. https://doi.org/10.2196/iproc.8441

[34] Tian Hao, Kimberly N Walter, Marion J Ball, Hung-yang Chang, Si Sun, and Xinxin Zhu. 2017. StressHacker: Towards Practical Stress Monitoring in the Wild with Smartwatches. In *AMIA Annual Symposium*. Washington D.C.

[35] J A Healey and R W Picard. 2005. Detecting Stress During Real-World Driving Tasks Using Physiological Sensors. *IEEE Transactions on Intelligent Transportation Systems* 6, 2 (Jun. 2005), 156–166. https://doi.org/10.1109/TITS.2005.848368

[36] Dirk H Hellhammer, Stefan Wüst, and Brigitte M Kudielka. 2009. Salivary cortisol as a biomarker in stress research. *Psychoneuroendocrinology* 34, 2 (2009), 163–171.

[37] Javier Hernandez, Rob R. Morris, and Rosalind W. Picard. 2011. *Call center stress recognition with person-specific models*. Vol. 6974. Springer, Berlin, Heidelberg. 125–134 pages. https://doi.org/10.1007/978-3-642-24600-5_16

[38] Karen Hovsepian, Mustafa Al'Absi, Emre Ertin, Thomas Kamarck, Motohiro Nakajima, and Santosh Kumar. 2015. cStress: Towards a Gold Standard for Continuous Stress Assessment in the Mobile Environment. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp) (UbiComp '15)*. ACM, 493–504. https://doi.org/10.1145/2750858.2807526

[39] 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 (Sept. 2019), 22 pages. https://doi.org/10.1145/3351249

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

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

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

[43] MC10. 2019. BiostampRC. https://www.mc10inc.com/our-products/biostamprc. [Online; accessed 06-May-2020].

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

[45] Varun Mishra, Tian Hao, Si Sun, Kimberly N. Walter, Marion J. Ball, Ching-Hua Chen, and Xinxin Zhu. 2018. Investigating the Role of Context in Perceived Stress Detection in the Wild. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers - UbiComp '18*. ACM Press, New York, New York, USA, 1708–1716. https://doi.org/10.1145/3267305.3267537

[46] Varun Mishra, Gunnar Pope, Sarah Lord, Stephanie Lewia, Byron Lowens, Kelly Caine, Sougata Sen, Ryan Halter, and David Kotz. 2018. The Case for a Commodity Hardware Solution for Stress Detection. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers - UbiComp '18*. ACM Press, 1717–1728. https://doi.org/10.1145/3267305.3267538

[47] Varun Mishra, Gunnar Pope, Sarah Lord, Stephanie Lewia, Byron Lowens, Kelly Caine, Sougata Sen, Ryan Halter, and David Kotz. 2020. Continuous Detection of Physiological Stress with Commodity Hardware. *ACM Transactions on Computing for Healthcare* 1, 2 (Apr. 2020), 1–30. https://doi.org/10.1145/3361562

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

[49] Eduardo Ogasawara, Leonardo C Martinez, Daniel De Oliveira, Geraldo Zimbrão, Gisele L Pappa, and Marta Mattoso. 2010. Adaptive normalization: A novel data normalization approach for non-stationary time series. In *The 2010 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[50] Nikolaos Passalis, Anastasios Tefas, Juho Kanniainen, Moncef Gabbouj, and Alexandros Iosifidis. 2019. Deep Adaptive Input Normalization for Price Forecasting using Limit Order Book Data. *arXiv preprint arXiv:1902.07892* (2019).

[51] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[52] Kurt Plarre, Andrew Raij, Syed Monowar Hossain, Amin Ahsan Ali, Motohiro Nakajima, Mustafa Absi, Emre Ertin, Thomas Kamarck, Santosh Kumar, Marcia Scott, Daniel Siewiorek, Asim Smailagic, and Lorentz E Wittmers. 2011. Continuous Inference of Psychological Stress from Sensory Measurements Collected in the Natural Environment. In *Proceedings of the 10th ACM/IEEE International Conference on Information Processing in Sensor Networks*. 97–108. http://ieeexplore.ieee.org/abstract/document/5779068/

[53] John C. Platt. 1999. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In *ADVANCES IN LARGE MARGIN CLASSIFIERS*. MIT Press, 61–74.

[54] Hans E Plesser. 2018. Reproducibility vs. replicability: a brief history of a confused terminology. *Frontiers in neuroinformatics* 11 (2018), 76.

[55] Polar. 2017. Polar H7. https://support.polar.com/us-en/support/H7_heart_rate_sensor. [Online; accessed 04-August-2020].

[56] Gunnar C Pope, Varun Mishra, Stephanie Lewia, Byron Lowens, David Kotz, Sarah Lord, and Ryan Halter. 2018. An Ultra-Low Resource Wearable EDA Sensor Using Wavelet Compression. In *Proceedings of the IEEE Conference on Body Sensor Networks (BSN)*. 193–196. https://doi.org/10.1109/BSN.2018.8329691

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

[58] Ensar Arif Sağbaş, Serdar Korukoglu, and Serkan Balli. 2020. Stress Detection via Keyboard Typing Behaviors by Using Smartphone Sensors and Machine Learning Techniques. *Journal of medical systems* 44, 4 (Feb. 2020), 68. https://doi.org/10.1007/s10916-020-1530-z

[59] Nazir Saleheen, Amin Ahsan Ali, Syed Monowar Hossain, Hillol Sarker, Soujanya Chatterjee, Benjamin Marlin, Emre Ertin, Mustafa Al'Absi, and Santosh Kumar. 2015. PuffMarker: A multi-sensor approach for pinpointing the timing of first lapse in smoking cessation. In *UbiComp 2015 - Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. Association for Computing Machinery, Inc, New York, New York, USA, 999–1010. https://doi.org/10.1145/2750858.2806897

[60] Akane Sano and Rosalind W Picard. 2013. Stress Recognition Using Wearable Sensors and Mobile Phones. In *Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, 671–676. https://doi.org/10.1109/ACII.2013.117

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

[62] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. 2018. Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction (ICMI '18)*. ACM, New York, NY, USA, 400–408. https://doi.org/10.1145/3242969.3242985

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

[64] Fred Shaffer, Rollin McCraty, and Christopher L Zerr. 2014. A healthy heart is not a metronome: an integrative review of the heart's anatomy and heart rate variability. *Frontiers in psychology* 5 (2014), 1040.

[65] Fernando Silveira, Brian Eriksson, Anmol Sheth, and Adam Sheppard. 2013. Predicting Audience Responses to Movie Content from Electro-Dermal Activity Signals. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '13)*. Association for Computing Machinery, New York, NY, USA, 707–716. https://doi.org/10.1145/2493432.2493508

[66] Feng-Tso Sun, Cynthia Kuo, Heng-Tze Cheng, Senaka Buthpitiya, Patricia Collins, and Martin Griss. 2012. Activity-aware Mental Stress Detection Using Physiological Sensors. In *Mobile Computing, Applications, and Services*, Vol. 76. 1–20. https://doi.org/10.1007/978-3-642-29336-8_16

[67] Terumi Umematsu, Akane Sano, Sara Taylor, and Rosalind W. Picard. 2019. Improving students' daily life stress forecasting using lstm neural networks. In *2019 IEEE EMBS International Conference on Biomedical and Health Informatics, BHI 2019 - Proceedings*. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/BHI.2019.8834624

[68] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T Campbell. 2014. StudentLife: Assessing Mental Health, Academic Performance and Behavioral Trends of College Students Using Smartphones. In *Proceedings of the International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*. ACM. https://doi.org/10.1145/2632048.2632054