This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2022.3188006

Toward Robust Stress Prediction in the Age of Wearables: Modeling Perceived Stress in a Longitudinal Study with Information Workers

Brandon M. Booth, Hana Vrzakova, Stephen M. Mattingly,
Gonzalo J. Martinez, Louis Faust, and Sidney K. D'Mello

Abstract—Given the widespread adverse outcomes of stress – exacerbated by the current pandemic – wearable sensing provides unique opportunities for automated stress tracking to inform well-being interventions. However, its success in the wild and at scale depends on the robustness and validity of automated stress inference, which is limited in current systems. In this work, we enumerate the properties of robustness and validity necessary for achieving viable automated stress inference using wearable sensors, and we underscore present challenges to constructing and evaluating these systems. Using these criteria as guiding principles, we present automated stress inference results from a large (N=606) *in situ* longitudinal wearable and contextual sensing study of information workers. Using a multimodal approach encompassing a wearable sensor, relative location tracking, smartphone usage, and environmental sensing, we trained regression models to predict daily self-reported perceived stress in a participant-independent fashion. Our models significantly outperformed baseline variants with shuffled stress scores and were consistent with small-to-moderate effects. Our findings highlight the performance disparity between robust and valid approaches to automated perceived stress inference and current approaches and suggest that further performance gains might require additional sensing modalities and enhanced contextual awareness than existing approaches.

Index Terms—Daily stress, wearable sensors, in-situ studies, phone agents, machine learning

1 Introduction

Stress in the 21st century is rapidly becoming one of the largest contributors to health decline, depression, and mental diseases [1], [2], [3]. The growing inter-connectedness of the world's workforce, global-scale competitiveness for jobs, increasing prevalence of night-shift work, and job automation efforts, are examples of trends which are both directly and indirectly negatively impacting workforce stress [4], [5]. Current rates of global economic market expansion are exerting more time pressures on workers, and stress related to job security and performance in the information workforce (e.g., accountants, managers, scientists, engineers) is projected to increase [6]. Though moderate amounts of stress may have beneficial effects on overall well-being [7], persistent daily stress erodes health (e.g. [8]), contributing to many negative physical and mental health outcomes, including heart disease, diabetes, depression, anxiety and insomnia [9], [10], [11], [12], [13], [14], [15], [16], [17], [18].

Stress-reduction interventions, meditation, and routine therapy have proven effective tools for managing work stress and mitigating its long-term effects [19], but individuals in westernized cultures may find it difficult to seek professional help. In a 2006 poll of American employees, approximately 40% of workers experiencing high levels of stress felt comfortable mentioning it to their employers, and among those that did, only 4 in 10 were offered professional help when reporting stress [20]. If the stress levels of individuals in the workforce could be efficiently monitored in real time, interventions could be (anonymously) deployed to help people manage stress in the moment or to seek help at a future time. This level of tracking would demand careful consideration of ethical and social control concerns, and its deployment to the general population would need to be conditioned on strict and enforceable security and privacy regulations.

Nevertheless, wearable commercial physiological sensors represent a rapidly growing industry that offers a unique opportunity for real-time stress tracking. Current wearable devices, in conjunction with smartphones and companion apps, offer real-time passive monitoring of certain physiological and behavioral signals known to be indicative of stress levels in laboratory studies, for example heart rate, heart rate variability, step count, and sleep quality [21], [22]. These sensors thereby provide a lens through which an individual's health status, well-being, and stress can be tracked on a daily basis [23], [24], [25]. Wearable devices have been widely adopted amongst consumers, with around 1 in 4 Americans reporting the use of wearable accessories, and their use is projected to increase [26]. However, understanding the link between the non-medical-grade data these pervasive devices collect and stress is still under vigorous investigation (e.g., [27]).

A number of recent studies have investigated the daily stress detection capabilities of certain physiological and contextual signals captured from consumer-grade sensors (e.g. wearables, cell phone apps) to make inferences about future stress states [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38]. The bases for these approaches are established on the successes of controlled laboratory experiments where some signals (e.g.

The research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA Contract No 2017-17042800005. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

- *B. Booth and S. D'Mello are with the Institute of Cognitive Science, University of Colorado Boulder, Boulder, CO, 80309.*
- *H. Vrzakova is with the University of Eastern Finland.*
- *S. Mattingly, G. Martinez are with the University of Notre Dame, Notre Dame, IN 46556.*
- *L. Faust is with Mayo Clinic.*
  Corresponding E-mail: brandon.m.booth@gmail.com

heart rate, heart rate variability, sleep) captured using medical-grade sensors are shown to be predictive of stress induced by manipulation (e.g. the cold pressor task [39], [40] or the Trier social stress test [41], [42]), which have limited ecological validity. The transition of these successful lab experiments to studies conducted in natural settings report mixed results [34], [35], [38], especially when comparing physiological and behavioral data against self-reported stress (e.g. [43], for review see [44]). Some studies demonstrate that contextual awareness of the activities of individuals (e.g., at work, socializing, exercising) can substantially improve the accuracy of stress prediction in natural settings [34], [35].

Though these works vary in their approaches to capturing physiological signals and assessing stress, they are united in their efforts to build a practical real-time stress inference system. Progress towards this goal appears imminently promising when considering the reported successes of these works in aggregate (see Section 2.4 for details). However, as we discuss below, many of these reported successes address simplified versions of the stress inference problem and are not robust reflections of the predictability of stress levels for general workforce populations at scale and in the wild.

In this work we aim to study the link between physiology, context, and stress in a robust and generalizable fashion. In particular, the contributions in this work are:

1) We enumerate the properties of robust and valid perceived stress inference intended for use in natural settings
2) We examine several recent works on wearable stress in the wild through the lens of robustness and validity
3) We analyze a large data set (N=606) of daily stress levels among information workers within five cohorts across the US
4) We present daily stress prediction results for this data set from robustly constructed machine learning models

2 Background and Related Work

2.1 Detecting and Measuring Stress

2.1.1 What is Stress?

Stress is a complex physiological phenomenon and the term was first employed by Hans Selye [45] to describe the bodily reactions of mice to non-specific nocuous agents (e.g. cold exposure, acute injury, excessive exercise, drugs). The short-term effects first noted by Selye are physiological in nature (e.g., fat tissue decrease, enlarged adrenal glands) and theorized to encompass the "general alarm reactions" of an organism to help it adapt and respond to threatening situations. The long-term health implications of prolonged stress include heart and liver disease [12], [46], diabetes [47], depression [10], [13], [15], [16], anxiety [9], insomnia [14], and other symptoms [11], [17], [18], [46]. Since then, our understanding of stress and its impacts has broadened. Stress can be acute or chronic, physical or psychological, and each type results in different impacts on the body depending on an individual's stress sensitivity.

2.1.2 Perceived Stress vs. Physiological Stress

Though physiological stress itself induces long-term biological maleffects, it is often linked with one's awareness and subjective perception of being stressed [48], [49]. Research has demonstrated a small-to-medium effect size between subjective self-reports of perceived stress and objective measures of physiological stress based on linear models and correlation analysis (see Section 2.1.3 for more details) [50], [51]. Perceived stress plays a unique role in the long-term impact of stress, and its relationship to physiological stress is not fully understood. Ven Eck et al. have noted that negative affect mediates the relationship between perception of stress during stressful events and physiological stress [52]. The converse has also been observed where high levels of perceived stress combined with a situational threat to one's coping ability may elicit negative affect [53]. Similarly, perceived stress has been observed to mediate the relationship between mindfulness and negative affect [54]. Furthermore, one laboratory study suggests that physiological stress is associated with subjective stress *only* if it is assessed during a stressor; physiological responses before or after a stressor are not associated with self-reported stress [55]. Thus, perceived stress seems to be both caused by and a cause for physiological stress, and it may provide a more holistic (i.e. aggregate) view of physiological stress over time.

2.1.3 How Can Stress be Measured

Regardless of whether stress is induced physically or mentally, it results in physiological changes which can be measured in a variety of ways. Biochemical markers such as cortisol, salivary α-amylase, plasma or urinary norepinephrine and its spillover rate, and interleukins each serve a role in the measurement of different types of acute or chronic stress [56], [57]. Stress also results in changes to the sympathetic nervous system (SNS) and parasympathetic nervous system (PNS), which can be detected using physiological sensing. Heart rate variability (HRV), for example, is known to differentiate between PNS and SNS activity [49], [56], [58], [59]. Other measurable physiological indicators include: electrodermal activity [60], heart rate and complexity [61], blood pressure [62], pupil size [63], and sympathetic nerve activity [64].

Changes to one's surroundings (e.g., one's social life, home environment, workplace conditions) can result in acute stress which is moderated by the context surrounding these changes. An individual's stress sensitivity will dictate how strongly these changes influence physiological stress, but it may influence perceived stress differently. Self-reports of stress obtained, for example, using the Perceived Stress Scale (PSS) [65] provide a subjective measure of perceived stress, which may contain information about both physical and mental stressors. Furthermore, other indirect measures of changes to one's surroundings or context have been used to gain insight into the relationship between perceived stress and physiological stress, for example social isolation [66], job-specific stressors [67], and location [68].

2.1.4 Stress Measurement in the Wild

Measuring stress in the wild, where a study's ecological validity can greatly improve the generalizability of its results, is challenging. Minimizing the burden of participation on study subjects is an essential concern when designing these types of *in situ* studies because it can directly influence compliance, attrition rates, and data quality [69]. Under these constraints, a continuous and passively collected measurement of stress is highly desirable.

Among the more common means of objectively measuring physiological stress is the cortisol test. Cortisol is produced when the ANS issues a fight-or-flight response to either acute or chronic stressors and serves as a reliable proxy for physiological stress in laboratory studies. However, these tests take anywhere from 1-5 minutes to perform and face a host of other challenges [70], so obtaining frequent measurements may threaten ecological validity. Other biochemical tests produce a similar participant burden, requiring brief and frequent interruptions in order to successfully capture stress dynamics throughout the day, which may also affect the validity.

Physiological measures of stress can be gathered continuously and have been proven effective for stress assessment

in the lab, especially HRV. Specific measures of HRV (e.g., standard deviation of cardiac cycle intervals, band-limited spectral power) collected from medical-grade devices have been useful in identifying different stress responses of the nervous system. Consumer-grade electrocardiogram sensors capable of obtaining these HRV measures are available but may be cost-prohibitive or uncomfortable for participants in longitudinal studies [69]. An increasing variety of other wearable consumer sensing technologies offer a promising avenue for continuously capturing stress measures and are increasingly being utilized in research studies (e.g., [27], [68], [71]). Some of these wearables incorporate multiple sensor capabilities such as photoplethysmography for detecting heart rate and peak-to-peak HRV or skin conductance sensors which capture galvanic skin response measures [33], [72], [73]. These devices are not as sophisticated as medical-grade sensors and thus do not offer the same amounts, types, or qualities of data as would be desirable, but they have still proven useful for capturing certain types of stress both in and out of the lab [31], [72].

As beneficial as physiological measures can be for stress assessment in the lab, studies have shown that they do not tell us the whole story [74], [75]. Some studies report that contextual awareness of an individual's activities can help boost stress prediction in the wild [34], [35], [36], [38]. One possible explanation is that physiological indicators cannot distinguish between different states of arousal, for example stress and excitement, therefore more contextual information is necessary to distinguish stress states. Since perceived stress is judged subjectively, it serves as a noisy filter admitting acute and chronic stressors and rejecting other states and forms of physiological arousal. Other potentially discerning information can be gleaned by capturing as much physical, cognitive, contextual, and behavioral information about individuals as possible. Modern portable and mobile sensing technologies can facilitate continuous collection of this range of information passively without interfering with participants during *in situ* data collection [69] and offer considerable promise.

2.2 Requirements for Robust and Valid Stress Detection at Scale in the Wild

Two highly desirable qualities in a daily stress prediction system are robustness and validity. A robust model has predictable outputs for a variety of inputs and is tolerant of data errors. Valid stress inference systems output accurate stress predictions for a large portion of the population. Facets of these qualities are discussed in detail below.

2.2.1 Robustness

Reliability: The accuracy of a model's stress prediction on unseen data (e.g. future data samples) should be similar for each new set of samples. Estimates of a model's reliability can be gathered, for instance, by measuring the variance of the distribution of accuracies on new data or during cross-validation, or by computing test-retest reliability metrics [76].

Missing Data: Data gathered in the wild is unavoidably faulty. Motion artifacts often plague the quality of data collected from wearables while separately, individuals may forget to wear devices, forget to clean them properly, allow the batteries to die, or accidentally break the devices. Therefore, missing, low quality, and corrupted data are inevitable. Ideally, a robust system would only make stress assessments with confidence after observing sufficient data, but in practice, this is not always possible due to the frequency of low quality and missing data. A practical robust system must make a best-guess estimation based on whatever information is or was previously available.

2.2.2 Validity

Generalizability: Data samples collected and used for model training need to be representative of the population(s) of interest so that the model is unlikely to be asked to make predictions on future novel samples. This implies that all of the usable data needs to be utilized during model training to ensure the model is exposed to a wider variety of sample data. Furthermore, it is imperative that the learning model is trained in a manner that does not perpetuate or create biases that lead to unfair stress predictions for certain subgroups of the population (e.g., based on gender or age). Nested participant-independent cross-validation is an indispensable tool for constructing and validating a generalizable model. A full discussion of methods for mitigating bias and preserving fairness in machine learning is outside the scope of this paper (see [77] for a review), but some example techniques include group blindness estimation, predictive group parity, and post-hoc adjustments.

Sufficient Accuracy: A robust daily stress prediction model for use in the wild should be capable of making predictions which are more accurate than a suitable baseline algorithm. For continuous-valued stress scores (e.g., from the Perceived Stress Scale [65]), a trained model should output better predictions than, for instance, a baseline model outputting the mean stress level across all subjects in the training data. The choice of baseline models should represent apples-to-apples comparisons. So, if a daily stress prediction model is personalized to individuals, then a sensible and fair baseline for comparison might be one which outputs each subject's expected mean stress level individually.

Convergent Validity: A model's stress predictions should correlate with ground truth stress so that the predictions can be used in their place. Stress scores obtained from psychometrically validated surveys, such as the PSS, have been tested for construct validity and can serve as appropriate ground-truth measures. Models which are designed to output predictions of these scores and which are also accurate should achieve convergence with these measures. Problems concerning this type of validity can arise when stress scores are transformed (e.g., via binarization) prior to model training, which can reduce the correspondence of the stress predictions to the original measured stress levels.

2.3 Prior Studies and Stress Prediction in the Wild

Recent daily stress modeling research conducted in natural environments spans a variety of populations, including students [33], [68], [71], [78], [79], information workers [30], [80], [81], and patients in hospital settings (e.g., pre-operative patients [82], pregnant mothers [83], elderly [84]). As with any study involving subjective human data, these studies face challenges unique to their respective populations and contexts in addition to common and typical challenges such as sample size, study duration, signal quality, and analytical methods [72]. We provide a summary table in the supplementary materials (see Table S1) categorizing recent *in situ* stress assessment research efforts according to their contexts, stress measurement tools and ground truth methods, number of subjects, study duration, signals and sensors, analytical techniques, and reported model performances.

Recent automated stress prediction efforts are united in their aims to infer daily stress from physiological, behavioral, and/or contextual information, but each one has its own unique combination of approaches, protocols, and analytical techniques. These differences make it difficult to directly compare the studies, but it does give us a greater sense of the range of

applications and potential impact of this line of research. A few general observations about these works in aggregate stand out. First, the link between measured stress and the signals from wearables and phone agents is fairly weak, as reported by prior statistical stress analyses [30], [80], [84], [88], [89]. This seems to contradict the high stress prediction accuracies achieved using ML methods (e.g., [22], [27], [33], [68], [78], [79], [82], [83], [85], [86], [87]). Second, this dichotomy may be explained by the other major difference between these two types of works, which is how stress is categorized prior to analysis. The papers focused on statistical analyses tend to preserve the stress values obtained from scoring their respective stress surveys (e.g., PSS) while the stress prediction studies typically discretize the stress values into a small number of categories. It appears that the best reported performances tend to arise from works using fewer stress classes (e.g., using binarization instead of trinarization), which suggests that this type of problem manipulation inclines towards over-simplification.

2.4 Analytical Challenges for Daily Stress Prediction in Natural Settings

Table 1 presents a list of challenges derived from our survey (Table S1) of *in situ* daily stress prediction research using wearable sensing technologies. This table focuses on methodological and AI-related challenges rather than data collection and experimental challenges (e.g., reducing participant attrition, increasing daily count of stress labels). Each row represents a unique challenge and contains a brief description of the challenge itself and the reason why it complicates stress research in natural settings. Some example references of this type of research are also provided for each row which illustrate the challenge. The stress prediction challenges are grouped by criteria according to Section 2.2 and categorized as being primary concerns for either robustness (rows 1-3) or validity (rows 4-9).

The challenges in the top portion of the table pertaining to model robustness (rows 1-3) are related to the lack of standard analytical procedures for research in this domain. Stress is person- and context-dependent, so the availability and quality of wearable sensor data in any particular study may strongly dictate the analytical process. Ideally, each analytical decision involving data processing and model training would have a standard and prescriptive set of best-practices appropriate for the application domain. However, for decisions involving data quantization (row 1), partitioning, and balancing (row 2) there are few established analytical norms for handling missing data (row 3) or ensuring reliable performance on future data samples in this stress domain. In the absence of a normative methodology, it is difficult for the research community to form a consensus and ascertain which types of approaches consistently perform well in new studies.

The lack of standard analytical approaches may be a symptom of the lack of a common set of tools for stress assessment, making it difficult to compare models and stress prediction results. In our survey of recent *in situ* daily stress modeling research, for example, stress is measured using the Perceived Stress Scale (PSS) [65], State-Trait Anxiety Inventory (STAI) [90], stress diaries [84], and custom Likert scale surveys (e.g., [72]) collected either retrospectively or via ecological momentary assessments (EMAs) [91]. Sometimes stress is not even measured, but asserted, depending on a participant's engagement in certain activities (e.g., [27]). A healthy variety of novel methods facilitates exploration, but it is difficult at present to ground existing research because analytical procedures are not unified across studies.

Other challenges pertaining to model validity are related to decisions made during the machine learning process. A traditional machine learning pipeline involves numerous stages of data processing to prepare for model training, including artifact removal, missing data imputation and/or exclusion, ground truth label generation, and feature extraction. At each point in this process, data processing may result in the incidental introduction of bias in the form of noisy data or label distortions. The effects of these distortions are accumulated throughout the machine learning pipeline and may result in reduced generalizability. For example, the partitioning strategy employed for model training, tuning, and testing can incidentally bias the trained model, artificially inflating its accuracy and reducing its generalizability. Using a partitioning strategy where each participant's data is completely contained in either the training or test set (not both) when performing cross-validation can help ensure the generalizability of the results and yield a realistic out-of-sample accuracy metric (row 4).

Another common theme we observed in our literature survey was the decision to simplify the daily stress prediction difficulty by either quantizing the daily stress labels or focusing on prediction of the extreme values while ignoring the mid-range stress labels, which are ostensibly the most difficult ones (row 5). For a problem as difficult as stress prediction in the wild, these types of simplifications are instrumental for developing an intuition and understanding of the limitations of stress modeling. In aggregate, however, the abundance of research efforts using these techniques and reporting high prediction accuracies may give an inflated impression of the state of stress prediction in the wild. In reality, these large performance scores are often representative of a subset of the sample population (e.g. people in the top and bottom 20% of reported stress levels) and not suitable for generalization to the whole population.

These potential sources of data bias can have a substantial impact on the resulting stress prediction performance. When comparing results to baseline models or results in other works, it is important to assess the statistical significance of the improvement of competing models (row 6). Apparent improvement in performance may be due to sampling noise or data biases, and measuring statistical significance is an instrumental tool for reducing the chance that reported performance gains are actually unrelated to a model's genuine improvement.

Additional challenges to model validity appear at the evaluation stage when measuring and reporting a trained model's performance. Quantization of the stress scores obtained from a validated stress survey reduces the amount of relevant information present in the emerging ground truth categories (row 7) and thereby reduces the overall correspondence of the stress predictions to the original stress scores (row 8). This is especially true when treating the stress scores as ordinal data as previous works have emphasized [92].

Finally, with all of these other factors potentially impacting the validity of a stress model, it is important for studies to report performance measures relative to reasonable baseline models (row 9). This allows models to be evaluated with respect to the unique contextual features of the domain of each study (e.g. student stress, hospital worker stress).

2.5 Novelty of Current Study

Our survey of *in situ* daily stress prediction research studies using physiological and contextual signals suggests too many differences and inconsistencies exist in the approaches, methods, and results to be able to assess how accurately stress inference can be performed in the wild. To establish a foundation

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2022.3188006

TABLE 1
Challenges to building robust and valid daily stress inference systems for use at scale and in natural settings

	Row	Criteria	Challenges	Rationale	Example References
Robustness	1	Reliability	Quantization thresholds are not standardized and vary across studies	Models using stress bins may disagree about stress levels	[78], [82], [85], [88]
	2		Imbalanced classes are balanced prior to machine learning	Trained models have a false notion of true stress distributions	[78], [82]
	3	Missing Data	Lack of common method for excluding a participant's data based on its available quantity	Hard to compare different models, and they may exclude different types of data	[68], [78]
Validity	4	Generalizability	Introduction of data or prediction bias (e.g., models are evaluated using sample-level k-fold cross-validation instead of subject-independent cross-validation)	Prediction performance on an individual participant or groups is uncertain	[22], [68], [79], [85]
	5		Quantized stress data in mid-range is omitted from analysis	All valid data should be included	[33], [68], [79], [88]
	6	Relative Accuracy	Performance of competing models is not tested for statistical significance	Unclear which model(s) perform the best	[33], [68], [85], [87]
	7		Continuous stress scores are quantized before modeling	Potentially relevant information about stress levels is lost	[81], [85]
	8	Convergent Validity	Quantized (e.g. binarized) stress may no longer correspond to true mental state	Predicted stress has diminished construct validity	[33], [68], [79], [85], [27]
	9		Performance of models is not tested against suitable baselines or potentially mediating factors	Unclear how to assess a model's performance relative to the data	[27], [72]

for robust *in situ* prediction of perceived stress from mobile sensors, this work presents results from a large-scale study of individuals in the information workforce. Several factors make this study unique. First, it includes data from a large (N=606) number of individuals, spanning five distinct cohorts, in their respective natural work environments. Second, the analysis focuses on maximizing the trained models' robustness and validity by addressing the challenges presented in Table 1. In particular, our study produces a robust model of stress prediction which avoids excessive simplification during data processing (e.g., quantization thresholds), predicts all stress scores from all participants, and employs a variety of baselines, models, and evaluation metrics. Our analytical approach seeks to maximize generalizability and construct validity by using subject-independent data partitioning, directly predicting stress scores, and grounding the model's performance relative to suitable baseline algorithms. The diverse participant pool, spanning multiple industries and information workforce jobs, used to train the model further enhances its generalizability. In short, our results provide a practical, robust, and valid benchmark for *in situ* perceived stress prediction of information workers from their wearable sensor data.

3 Methods

3.1 Data Set Description

This study was approved by the University of Notre Dame's IRB under protocol number 17-05-3870. A comprehensive explanation of the data collection procedures is available in [73]. Key components are summarized in the following subsections.

3.1.1 Participants

A total of 606 full-time, salaried information workers (e.g., consultants, engineers, business/finance workers who primarily work with data) were enrolled in a year-long observational study starting in mid-2018 (prior to the COVID-19 pandemic). Participants were recruited from across the U.S. from employers and from community message boards. Four cohorts were established from four different institutions: a nationally distributed tech services firm (cohort 2), a large mid-western United States tech and engineering firm (cohort 3), a small mid-west United States software firm (cohort 4), and a medium-sized mid-western university (cohort 5). A fifth cohort (cohort 1) was obtained from interested applicants from assorted channels unaffiliated with the other institutions (e.g. friends of recruited participants, respondents to newspaper articles). The participant pool included a range of genders, occupations, income levels, education levels, and job roles. Demographic statistics for these participants are presented in Table S3.

3.1.2 Sensing Devices

The study utilized multiple sensing devices to capture a variety of physiological and behavioral information streams, providing time series data and snapshots of different aspects of daily life (i.e., physical activity, workplace behavior, phone usage, sleep, heart rate, weather, and day of the week). See Section 3.2.1 for an explanation of the rationale for capturing these signals.

Participants were provided a commercial-grade **Garmin Vivosmart 3** to collect heart and physical activity measures. This device obtained approximations of heart rate using photoplethysmography (PPG) [93] and assessed participant motion, step count, and physical activity using an accelerometer. Previous studies have established the accuracy of this device and

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2022.3188006

other types of PPG sensors by comparing its extracted heart rate to electrocardiogram measures from a chest strap [94] or by comparing its heart rate variability features to research-grade PPG sensors [23]. In aggregate, the correlation between commercial-grade PPG and classic ECG HR and HRV features is excellent and appears robust to differences in skin color [23], but it degrades as physical activity increases [95]. In our own experiments, 136 volunteers among our participants elected to wear both a Zephyr chest strap (ECG) and Garmin Vivosmart 3 (PPG) during the enrollment period, and 21 of them had sufficient HR confidence (above 20%, computed by Zephyr) for a period of at least 10 minutes. We compared the beat-to-beat intervals of the two sensors over these periods of time (mean duration of 26.2 minutes) using sliding time windows of varying lengths (3s to 300s) and observed Pearson correlations between 0.7 and 0.79. Garmin has also published results showing that sleep stage timing and duration measurement is about 70% accurate in real-world conditions [96].

A custom smartphone application (**Phone Agent**) was installed on participants phones to collect phone usage metrics such as number of phone unlocks, active screen time, and GPS location [97]. Though the raw GPS was captured and recorded, it was only used to extract coarse location data from Foursquare, a location tracking app, and to estimate the weather. Participants were also provided **Bluetooth beacons** to be placed in their offices and homes. The signals from these beacons were detected and timestamped by the smartphone application, providing a proximity-based measure of when participants were at home or in the office. Additionally, participants received smaller, key-fob beacons which could be kept on their keychains or in wallets/purses and which captured periods when two participants were in close proximity. Additional **contextual information** about the weather was collected daily for participants within range of their Bluetooth beacons at home, using the zip code of the beacon and a web interface available from the National Oceanic and Atmospheric Administration.

3.1.3 Data Collection Protocol

All participants provided written informed consent prior to taking part in the study. Participants first completed a set of individual differences questionnaires (not analyzed here), upon which they were affixed with the devices. They were instructed to wear the Garmin Vivosmart 3 at all times starting on the day they completed enrollment, excluding time spent charging the device and taking showers. Constant wear allowed the device to capture daily measures of a participant's step count, heart rate and variability, and measures of sleep quality including sleep phases, time to bed, and awake time. Battery life at full charge lasted approximately 5 days and the device could be charged to full in as little as half an hour. Researchers recommended the device be charged any time a participant was showering. Participants were able to report lost or broken devices and receive replacement devices at any time over the course of the year-long study.

Participants placed battery-powered Bluetooth beacons at home and at work desks. The beacons could function for up to 18 months while requiring no participant intervention for the duration of the study. Participants were instructed to keep the key-fob beacons on their person at all times so their proximity to the Bluetooth beacons could be measured. The battery level of these key-fobs was tracked during smartphone sightings, and participants were periodically sent reminders to change batteries when low. The smartphone app tracked these sightings and also recorded and reported various metadata including the time since last check in with the server and the up-time of the app.

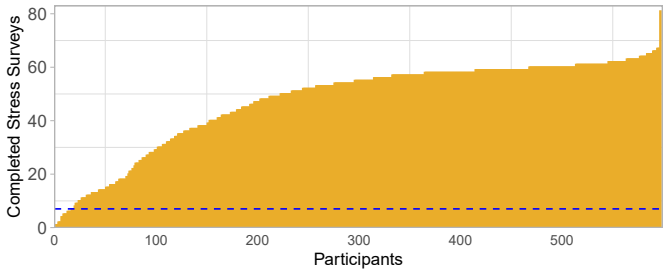

Fig. 1. A bar graph showing the number of completed daily stress surveys for each of N=606 participants. The horizontal dashed line denotes our threshold of survey compliance for inclusion in the modeling process (# surveys $\geq$ 7).

3.1.4 Perceived Stress Assessment

For the first 56 days following enrollment, daily SMS messages were sent to the participants' phones containing links to surveys asking them to record their momentary level of stress and also answer questions about their health, mood, and sleep [73]. The surveys were sent at either 8am, noon, or 4pm following a semi-random schedule, and participants' responses were accepted up to 4 hours later. Perceived stress was assessed using the following question on a 5-point Likert scale: "Overall, how would you rate your current level of stress?" (Response scale: 1=No stress at all, 2=Very little stress, 3=Some stress, 4=A lot of stress, 5=A great deal of stress). This stress item was validated by MITRE Corp. in an unpublished study involving 991 crowd-sourced participants, and the correlations with this item and other validated stress measures are tabulated in Table S2 in the supplemental material. The perceived stress measure correlations were moderate and positive for state and trait anxiety, negative affect, and neuroticism (.51<r<.61, demonstrating convergence), but were negative for positive affect (r = -.33, demonstrating discrimination), on par with results from validation studies of 10-item perceived stress scales [98], [99].

3.1.5 Inclusion Criteria

In order to minimize the impact of severely insufficient stress reports on modeling efforts, we required 7 of the roughly 56 daily stress surveys (see Section 3.1.3) to be completed per participant. We chose 7 with the aim of excluding participants who did not provide at least one week's worth of data. Among the 606 total participants available to the research team, 597 of them met this requirement and were used for further analysis. Figure 1 illustrates the number of daily stress surveys completed by each participant with a dotted line denoting our compliance threshold. These 597 participants produced a total of 28,226 samples of daily stress survey scores and associated data. Overall, the stress reports were highly skewed towards the lower end of the scale with fewer than 5% being greater than three (see Figure 3).

3.2 Constructing a Perceived Stress Inference System

3.2.1 Feature Extraction

An assortment of features was extracted from each of the signals provided by the wearable sensors and corresponding to known correlates of stress based on prior literature. Table 2 provides a summary of the sensors used, signals obtained,

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2022.3188006

TABLE 2
A list of features extracted from signals obtained from our wearable sensors study

Sensor	Signals	Features
Garmin Vivosmart 3	heart pulse (PPG), motion, ambient light	HR, HRV (rmssd, sdnn, sdann), step count, illuminance, recent sleep (start, end, duration, rolling, sleep debt, phase duration)
Phone Agent	location, phone usage, ambient light, motion	distance traveled, call (in count, out count, duration, missed count), unlocks (count, duration) activity (on foot, biking, driving, tilting, sleeping), illuminance
Bluetooth Beacons and Key-fobs	location proximity	duration at location (at work, at home, in bed, commuting, at desk, away from desk), work activity (start time, end time, break count), social interactions (time spent near 0,1,2,3+ others)
Environment	weather, sunrise, sunset, time of day	temperature, precipitation, humidity, wind (speed, chill, feels like), visibility, pressure, cloud cover, heat index, snow fall, sunrise and sunset time

PPG = photoplethysmography, HR = heart rate, HRV = heart rate variability, rmssd = root mean square of successive RR interval differences, sdnn = standard deviation of NN intervals, sdann = standard deviation of 5-minute averaged NN intervals. Note: All features were aggregated per day. HRV features were aggregated over various windows of time relative to survey completion and work start/stop hours (e.g., 30 minutes prior, from 8am to 6pm). Phone agent features were additionally aggregated within each epoch, and the Garmin Vivosmart features were also aggregated within these windows: current time, daily, weekly, during work, not during work, 15min prior to start of work, hour at start of work, hour at end of work, and within each epoch period.

features extracted from the signals, and the windows of time over which the extracted features were based.

Heart rate variability and measures of physical activity (motion, activity type, and heart rate) have been linked with stress in a number of prior studies [49], [61], [100]. In the present study, these measures were extracted from the Garmin Vivosmart 3, which also provided the number of steps taken and time spent being physically active. Sleep duration and quality are also well-understood contributors to perceived stress and physical energy levels [101], [102]. The duration of the primary period of sleep activity (excluding short naps) and sleep quality metrics (e.g., rolling, sleep phase durations) were extracted from the Garmin device using its approximate timestamps for bedtime and awake time.

Personal smartphone usage has also recently been linked with stress, especially due to a rise in the number of push notifications, social media participation, and a growing expectation of responsiveness [103], [104], [105]. We captured measures of smartphone interactions while participants worked by counting the number of phone screen unlocks, screen-on durations, and the number and durations of phone calls. Furthermore, extended stays at either home or work (e.g., working late, working on weekends, never leaving home outside of work) have also been linked to sleep disturbances, exhaustion, and in some cases physical injuries [106]. The proximity of participants' smartphones to Bluetooth beacons and the GPS data were used to produce several location-based features, such as the number of work sessions, number of work breaks, time at the work desk, number of unique places visited, time spent in vehicle, and total distance traveled. Finally, because participants carried Bluetooth key-fobs in addition to having static beacons, mutual discovery allowed for the creation of social interaction features such as daily interaction quantity and duration.

Changes in environmental conditions based on season (e.g., day light, temperature) are a known stressor for some individuals with seasonal affective disorder [107], [108] but may also have indirect effects on stress, for example, due to increased traffic or effects on sleep duration [109], [110], [111], [112]. To capture stress effects resulting from weather changes, the average weather conditions, 24-hour time of sunrise and sunset, and temperature were recorded for each participant based on their GPS location when near the Bluetooth beacon at home.

3.2.2 Feature Aggregation

To facilitate the investigation of the dynamics of stress over the course of a day, the extracted features from all signals were aggregated in time. Table 2 lists the time windows considered for features derived from each sensor. All features were aggregated per day using a typical set of statistical functionals: mean, median, min, max, range, variance, and standard deviation. Some of the features extracted from the Garmin Vivosmart 3 were aggregated using the same functional set applied over other time windows as well, for example during work or while away from work. Physiological features and phone usage and activity features were aggregated using the same functions across different time spans during the day (epoch 1: 12am - 9am, epoch 2: 9am - 6pm, and epoch 3: 6pm - 12am) and also aggregated according to location context (e.g., at home, at work, in the car) and time relative to survey response (e.g., within 30 minutes or 1 hour prior to survey completion). All date-time and categorical features were converted to numerical values for machine learning. In total, there were 488 features with these feature functionals included (236 from the Garmin Vivosmart 3, 131 from the Phone Agent, 103 from Bluetooth beacons, and 18 from the environment).

3.2.3 Machine Learning Pipeline

Partitioning and Stratified Cross-validation: Five data folds were extracted from the 28,226 samples of daily participant data using a subject-independent stratified sampling method, described as follows. The daily stress reports were averaged per subject, leaving 597 mean stress values, each corresponding to a participant. A binned distribution of these mean stress scores was obtained using deciles. Five mutually exclusive groups of these scores were extracted by randomly sampling 20% of the scores from each of the ten bins without replacement, resulting in five sample groupings with equivalent binned distributions. The resulting five lists of participants formed the five subject-independent folds, which were used to train a stress inference system using nested cross-validation. Note that resampling techniques were not used to balance instances of the five stress labels (see Figure 3) to avoid giving the model a false sense of the uniformity of stress in the wild (i.e., row 2 in Table 1). Figure 2 gives an overview of the partitioning, training, and testing scheme.

Missing Data Imputation: Given the challenging nature of conducting *in situ* human sensing studies [69], some proportion

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2022.3188006

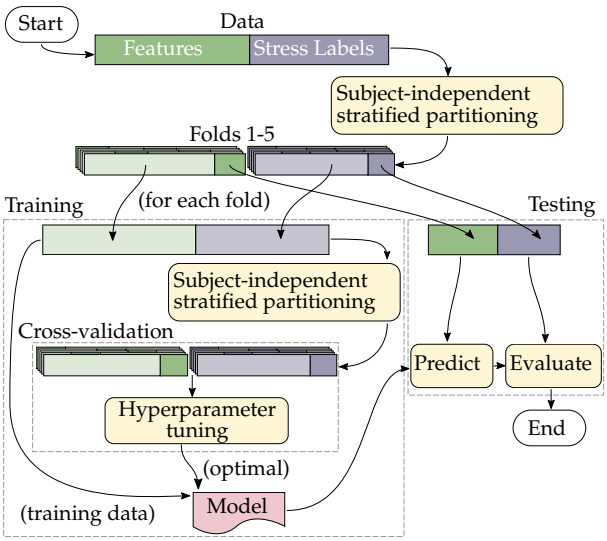

Fig. 2. An overview of the machine learning training and testing strategy using a subject-independent data partitioning scheme (best viewed in color).

of the observed data was non-compliant and missing due to factors such as dead sensor batteries, sensing failures, and participant attrition. We estimated 9.5% of all data collected was missing due to non-compliance by counting the number of missing features per day per participant where no valid data was present from any portable sensor (i.e., Garmin Vivosmart 3, phone agent, or Bluetooth beacons). Furthermore, because of the highly context-dependent nature of the features we extracted (e.g. number of beacon detections at work, HRV at the end of work, duration of phone calls), data for many daily features were missing due to the participants' daily schedules, actions, and choices on any particular day (e.g., not going into work on an off-day). On days where at least some valid participant data was recorded, approximately 32% of the features were irrelevant and missing due to these factors. Though in principle it would be interesting to investigate fully context-dependent modeling which only uses data from valid sources (e.g., only using work-related features on work days), we elect to conduct cross-subject analysis by imputing missing data. We tested several methods for imputing missing data within each feature, including zero-fill, mean-fill, person-specific mean-fill, and a more advanced *multiple imputation with denoising autoencoder* approach [113]. Although these imputation methods are conceptually unique, their impact on the resulting models' performance in our case was negligible. Therefore, all results reported in this work are derived from a mean-fill imputation method calculated and applied separately per fold during cross-validation.

Model Selection: Stress prediction was conducted using the full set of features previously described and a variety of machine learning models. Elastic net (EN) and random forest (RF) algorithms were selected for their interpretability and ability to reduce the effective feature set during training. A feed-forward multi-layer perceptron (MLP) was used as a baseline for deep learning approaches, and two time-aware methods including gated recurrent-unit (GRU) networks and long short-term memory (LSTM) networks were employed to predict stress based on the previous three days of data. These algorithms were tuned using a nested five-fold cross-validation strategy with participant-independent folds. In turn, each of the five precomputed data folds obtained from the stratified partitioning strategy (described earlier) were held-out and used for model evaluation. The remaining data (training data) in each iteration was further partitioned into five folds using the same participant-independent stratified sampling technique and used for hyperparameter tuning. Data imputation and z-normalization were fitted and applied independently to each feature in the training set during training and applied to the held-out portion for validation and testing. For each of the two learning algorithms, the hyperparameter set with the best average performance during tuning was used to retrain a model on all training data and make predictions on the held-out data (see Figure 2).

Model Tuning and Evaluation: For the RF and EN methods, separate models were trained for each of two final evaluation metrics: Spearman's correlation coefficient (ρ) and symmetric mean absolute percentage error (SMAPE) [114], both of which were also used for hyperparameter tuning. The particular formulation of the EN objective function was as follows:

$$\frac{1}{2n}\|s - Xw\|_2^2 + \alpha\beta\|w\|_1 + \frac{1}{2}\alpha(1-\beta)\|w\|_2^2$$

where s and X denote stress labels and the data respectively, w denotes the linear weights, and α and β are tunable constants. The hyperparameters for this model were tuned over a grid (Spearman: $\alpha \in \{0.01, \mathbf{0.1}, 1.0\}$, $\beta \in \{0.6, 0.7, 0.8, \mathbf{0.9}, 0.95, 0.99\}$; SMAPE: $\alpha \in \{10^{-30}, \mathbf{10^{-20}}, 10^{-18}, 10^{-14}\}$, $\beta \in \{10^{-40}, 10^{-20}, \mathbf{10^{-12}}, 10^{-6}\}$). Random forest parameters were also tuned over a grid (Spearman: number of trees $\{100, \mathbf{500}, 800, 1200\}$, maximum forest depth $\{10, \mathbf{20}, 50, 100, \infty\}$; SMAPE: number of trees $\{\mathbf{100}, 500, 800, 1200\}$, maximum forest depth $\{10, 20, 50, 100, \boldsymbol{\infty}\}$). For the neural methods, the number of layers (depth) and nodes per layer (width) in the MLP were tuned over a grid in a similar fashion to [38]: width $\in \{10, 20, \mathbf{30}, 40\}$, depth $\in \{1, 2, \mathbf{3}, 4\}$, and we also tested different loss functions (smooth L1 loss, **mean squared error**), optimizers (stochastic gradient descent, **adam**), activation functions (**Gaussian error linear units**, rectified linear units), and learning rates (10^{-1}, 10^{-2}, $\mathbf{10^{-3}}$). Based on results from early tests, we used a batch size of 32 to encourage adequate exploration of the loss function space and we trained over 50 epochs for computational tractability. The same neural network structure and learning parameters (but not weights) resulting from hyperparameter optimization of the MLP network was used for the GRU and LSTM models, except these models included an extra input layer of width $\in \{10, 20, \mathbf{30}, 40\}$ of either GRU or LSTM units respectively. The optimal hyperparameters for all models appear in bold font.

The pipeline was implemented using Python 3.6, Scikit-learn 0.20.2 [115], Tensor Flow 2.6.0 [116], PyTorch 1.10.1 [117]. Data is available at: https://tesserae.nd.edu/data-sharing/ and the modeling and data analysis code is available at: https://github.com/emotive-computing/mosaic_stress_2021.

3.2.4 Experiments

For each of the models, *standard*, *shuffled baseline*, and *within-subject shuffled baseline* experiments were conducted. The *standard* version followed the methodology described thus far where subject-independent folds and nested cross-validation were employed to predict stress. In the *shuffled baseline* variant, the daily stress scores were randomly shuffled across the entire data set prior to training. The *within-subject shuffled baseline* similarly shuffled the stress labels but ensured that the randomized mixing was performed within each subject's sample data. These

two shuffling algorithms provided a simulated noise baseline for comparison with the *standard* version.

4 Results

4.1 Perceived Stress Prediction

Figure 3 illustrates the distributions of self-reported (i.e., ground truth) and predicted stress scores across all surveys and all participants for the MLP model and the RF model optimized for Spearman correlation in the *standard* experiment. Instead of evaluating the models on instance-level stress predictions, we computed performance measures per participant and then aggregated these measures across participants to give a more realistic perspective on the expected performance of the models on future participants. Table 3 summarizes the distributions of participant-level performance metrics for each model in the *standard* and *shuffled baseline* variants (see Figure S1 for histograms).

Two-tailed paired-sample t-tests indicated that the *standard* EN, RF, and MLP models achieved significantly higher ($p < 0.001$) Spearman correlations than both *shuffled baseline* variants, but failed to significantly outperform the baselines using SMAPE. Both the GRU and LSTM models performed comparably in both Spearman correlation and SMAPE when compared to their shuffled baseline variants, indicating chance-level accuracy. If we compare the per-participant distributions of Spearman correlations in the *standard* experiment among the top three models using a two-tailed paired sample t-test (not shown in Table 3), we find that the RF model significantly outperformed the EN model while the mean difference between the RF and MLP models is not significant.

An ordinal evaluation metric, such as the Spearman correlation coefficient, captures the relative differences in stress levels on different days and therefore may be preferable to interval-scale metrics given the skewed distribution of stress scores (see Figure 3). Furthermore, since the output stress prediction range is constricted relative to the ground truth, Spearman correlation offers a more practical measure of model performance. For these reasons, we proceed with further analysis using Spearman correlation as the primary performance metric.

Both the RF and MLP models demonstrated a significant gain in Spearman correlation (upwards of 0.15) and performed comparably. Figure 4 shows a histogram of the within-participant Spearman correlations for the *standard* RF model, which achieved the highest mean performance, and shows that for 82% of participants, stress predictions of the RF model had higher Spearman correlations than the *within-subject shuffled baseline* mean (indicated by the dark blue line).

4.2 Comparison to Other Stress Models

Though not significantly different from the MLP, the RF model achieved the highest average Spearman correlation, so we selected this model for comparison to two other physiological measures of stress: the Garmin wristband sensor's stress score and HRV (inversely related to stress in stress elicitation lab studies [49]). Garmin's stress score is intended to be used for tracking well-being over time and is a public-facing but proprietary measure. The HRV measure was calculated from the beat-to-beat intervals (BBI) using the standard deviation of average normal-to-normal formula (SDANN; an international standard for long-term HRV [118]). Per-minute BBI values were obtained by averaging over sliding five-minute windows as long as data from four of the five minutes were present. The standard deviation was computed from these minute-level

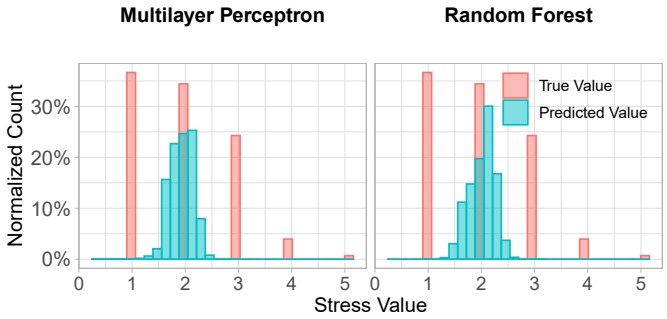

Fig. 3. Distribution of daily reports of stress levels across all participants and the stress level predictions from the standard MLP model and the RF model (optimized for Spearman correlation).

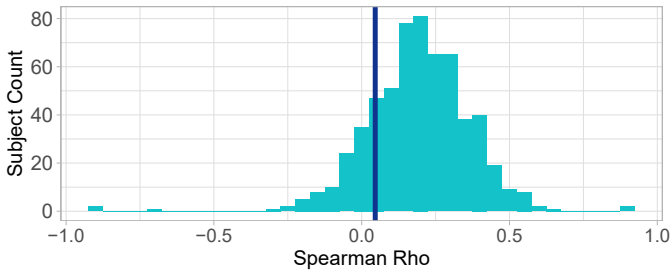

Fig. 4. Histogram of within-participant Spearman correlations computed using RF predictions and ground truth stress. The dark blue line marks the top-performing baseline model's mean (*shuffled within-subject*). The RF model performed better than the baseline for 82% of the participants.

TABLE 3
Summary statistics of the distribution of performance metrics applied to individual participants

	Experiment	Model	ρ (σ)	s (σ)
Baselines	Shuffled Cross-subject	EN	-.01 (.12)	.38 (.06)
		RF	-.01 (.15)	.39 (.06)
		MLP	-.00 (.18)	.39 (.06)
		GRU	-.00 (.18)	.43 (.06)
		LSTM	.00 (.17)	.45 (.06)
	Shuffled Within-subject	EN	-.00 (.12)	.38 (.06)
		RF	.05 (.18)	.39 (.05)
		MLP	-.01 (.18)	.39 (.13)
		GRU	-.00 (.18)	.46 (.12)
		LSTM	-.02 (.18)	.47 (.11)
Standard	Standard	EN	.16 (.16)*†	.39 (.05)
		RF	**.19 (.18)***†	.39 (.05)
		MLP	.18 (.19)*†	.39 (.12)
		GRU	-.00 (.17)	.47 (.08)
		LSTM	-.01 (.18)	.48 (.08)
		RF+STAI	**.25 (.19)***†	-
Other	Comparison	Garmin	.00 (.21)	-
		HRV	.01 (.28)	-

EN = elastic net, RF = random forest, MLP = multilayer perceptron, GRU = gated recurrent unit network, LSTM = long short-term memory network, STAI = state-trait anxiety inventory, ρ = Spearman correlation, s = SMAPE, σ = standard deviation of metric across participants, $*$,† = significant improvement in mean difference from *shuffled baselines* (*) or *other comparison* models (†) using a paired-sample two-tailed t-test ($p < 0.001$).

averages between 8am and 6pm, resulting in a daily HRV measure.

We used the Spearman correlation measure to assess the performance of each model for each participant because it is agnostic to range differences in each model's stress representation (i.e., it used ranks for comparison). The RF model achieved a mean Spearman correlation $\rho = 0.19$ while the Garmin stress model obtained $\rho = 0.00$ and the HRV model $\rho = 0.01$. A set of pairwise two-tailed paired-sample t-tests revealed the mean difference of about 0.18 between the RF and other models was significant ($p < 0.001$) in both cases while the 0.01 mean difference between the Garmin and HRV models was not significant ($p = 0.47$). Furthermore, we computed one-sample two-sided t-tests for each model to assess whether their per-subject Spearman correlation distributions could be generated by a process with a mean of zero. The RF model's ρ distribution was determined to be significantly distinct ($p < 0.001$) while neither the Garmin ($p = 0.92$) nor HRV ($p = 0.35$) models were significantly different, indicating chance performance. These results demonstrate that the RF model, using features derived from wearable sensor data, was able to model some variation in stress while the Garmin and HRV model results are indistinguishable from purely uncorrelated stress predictions.

4.3 Feature Information

The top 15 features selected by the two interpretable models (EN and RF) are shown in Figure 5. The top 15 EN features accounted for about 56% of the total feature weight, and due to the sparsity induced by the L1 regularization term, only 162 of the total 488 features received non-zero weights. The top 15 features for the RF model accounted for approximately 22% of the total weight of all features, and none of the features received zero weight. In both models, the top 15 features consist of a combination of information from all sensors (see Table 2) demonstrating that successful perceived stress inference is a multi-modal problem which relies on the union of contextual, physiological, and behavioral information. Furthermore, Figure 5 color-codes corresponding features that were separately selected among the top 15 by both models. The top choice in both models by a large margin was a binary indicator of whether a participant went to work on a particular day. The other top features shared by both models were humidity and the average difference in sleep duration between weekdays and weekends (suggesting recovery from sleep debt). These three shared top features may be more reliable predictors of stress than others due to their utility in both models. Work is the primary source of stress for a majority of information working professionals, so it is no surprise to see the "at work" feature at the top of both models. The shared humidity and weekend/weekday sleep difference duration features support findings in other works noting their importance for stress assessment [107], [108], [109], [110], [111].

4.4 Enhanced Model with State and Trait Anxiety

In a separate experiment, we tested adding a pertinent anxiety trait measure to the *standard* model to see if knowledge of individual differences in baseline anxiety levels would improve perceived stress inference, as many works have established a link between anxiety and stress (e.g., [119], [120], [121], [122]). The anxiety measures came from the *State-Trait Anxiety Inventory* (STAI) questionnaire [123] completed by participants during study enrollment and prior to the beginning of the study period. Figure S2 shows the distribution of STAI scores across participants, and readers are referred to [73] for more information about the pre-study survey. Borrowing the same optimal hyperparameters, we retrained the top-performing RF *standard* model with this STAI feature and observed an increase in Spearman correlation from 0.19 to 0.25. The relative difference in performance is significant ($p < 0.001$, two-tailed paired-sample t-test) when adding this single feature and results in a Cohen's d effect size increase of 0.13, indicating that individual differences in stress sensitivity hold additional pertinent information for perceived stress inference.

4.5 Effect of demographics and personal traits on stress prediction (Generalizability)

Personal and demographic traits may be associated with unique physiological signal patterns [124], [125] and thus may serve as moderating variables for daily stress scores. To systematically explore this hypothesis, we examined how individual differences in stress levels varied with respect to demographic factors: age, gender, language proficiency, supervision role, education, income, and cohort. We also controlled for the number of days participants were compliant. Table S3 shows standardized linear regression beta coefficients and p values between these factors and 1) averaged daily self-reported stress, 2) averaged RF stress predictions, and 3) Spearman correlations between the perceived stress and RF stress predictions. We found that in general demographics were not significant predictors of either perceived stress, predicted stress, or differences in stress predictability between participants, suggesting no moderation by demographics. Unsurprisingly, the more compliant participants reported lower stress scores and this was also detected by the RF model (*RF stress predictions* column). However, model accuracy (*RF Spearman* column) was not predicted by compliance.

4.6 Comparison to Other Stress Studies in the Wild

Many studies focused on understanding the link between contextual and physiological factors and perceived stress have reported model performance on binned data (recall rows 1, 7, and 8 in Table 1 regarding challenges to robustness and validity). Though this work aims to report performance measures for robust and valid modeling approaches, we include such binary classification metrics only for comparison to similar studies.

A common approach used to assess accuracy in stress prediction performance involves the training and testing of models on binarized stress labels (e.g., [34], [35], [37]). To mitigate per-participant binning variance due to individual differences in stress valuation, we computed the median stress score for each participant based on their responses and then split their scores into *low* and *high* bins. We tuned, trained, and evaluated three classification models to predict these binary labels using 5-fold nested cross-validation: a random forest, k-nearest neighbor classifier, and a support vector machine (see the Github code for details). The random forest yielded the highest F1-score of 0.75, which is similar to the F1-score of 0.77 reported by Mishra et al. [34] when both physiological and contextual features were used to predict perceived stress in the wild. Our accuracy (0.62), precision (0.65), and recall (0.89) values were also higher than those reported by Soto et al. [37] (0.54, 0.25, 0.44, respectively) for their models trained on physiological and computer interaction data in an 8-week study of 14 participants in the workplace.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2022.3188006

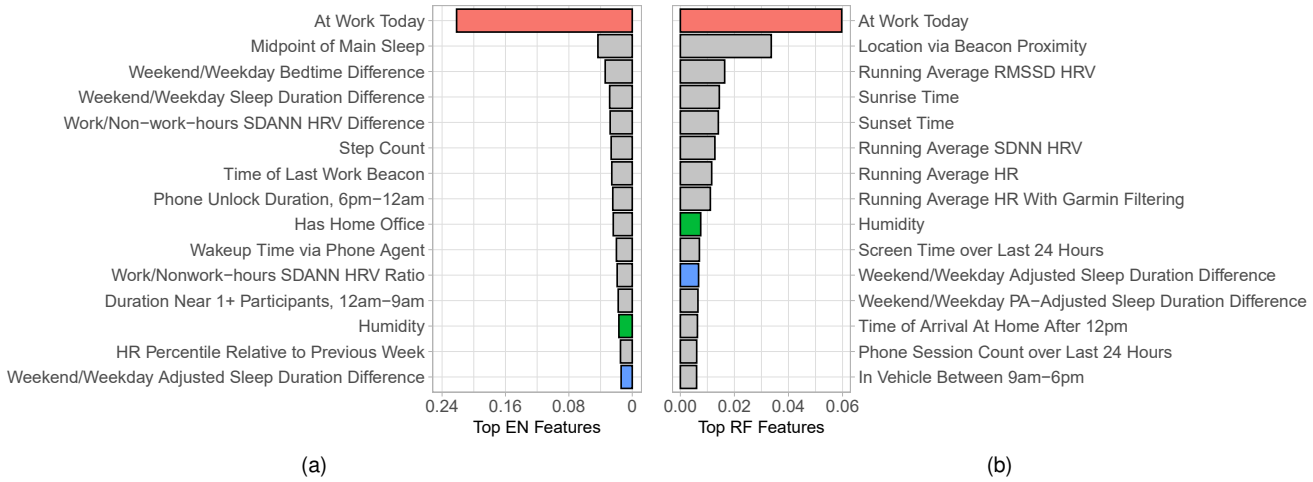

Fig. 5. Relative importance of the top 15 features in both the EN (left) and RF (right) models. Features appearing in both models are colored the same.

5 Discussion

Understanding the dynamics of perceived stress prediction in the wild for different individuals using contextual, environmental and physiological information is a difficult prospect but one which could greatly improve mental health and well-being if done properly. Our results, like many other studies, demonstrate that physiological and behavioral features capture meaningful information about stress, but it also points out that empirically these features only account for a modest amount of the observed variance. In the remainder of this section, we reexamine our main findings and consider limitations and future work.

5.1 Main Findings

The stress prediction results in Table 3 demonstrate that the best *standard* model (RF) offers some predictive power for perceived stress prediction when compared to the shuffled baselines. However, Figure 3 illustrates that this model does not capture the same magnitude of variability or range of stress responses present in the data. It should be noted that stress scores of 4 and 5 occur in less than 5% of the data, so the trained models may not have sufficient information to be able to predict these scores. The limited range of output values implies that the trained models are not able to distinguish between the sampled features corresponding to extreme levels of stress compared to those associated with more average stress values. This observation is in line with other recent *in situ* studies conducted in the wild (e.g., [72]).

Our top-performing RF model not only outperforms the shuffled baselines but also both the Garmin stress and HRV stress models. Using the ordinal Spearman metric, the two latter models appear insufficient for use in the wild as their performance is indistinguishable from an uncorrelated process. Stress prediction in the wild therefore seems to be highly multi-modal and demands an understanding of contextual factors beyond physiology, which has also been noted in several recent works (e.g., [34], [35]). The presence of sunrise time, weather, and other temporally non-local features among the most important features appearing in Figure 5b supports this as well.

We have made extensive efforts to ensure that our models were robustly trained according to the criteria laid out in Section 2.2. We established the robustness of our models by demonstrating the per-participant correlations were above the shuffled baseline performance for 82% of people in our study (see Figure 4) and that model with the highest mean correlation (RF) significantly outperformed the best baseline's mean performance ($\rho = 0.05$, RF shuffled within subject). Missing data was imputed without peeking into the test set while nearly all the available data was used for training and testing, and the model was able to achieve better stress prediction performance than the two shuffled baselines, Garmin stress, and HRV stress models.

Additionally, we have established the validity of our models through the data processing and machine learning pipeline procedures. We trained our models using subject-independent folds and cross-validation to help ensure generalizability. Each model was trained to measure performance on the perceived stress scores directly, avoiding unnecessary quantization and potential information loss. We also examined various demographic factors and showed that, apart from the number of days of compliance, there were no inadvertent demographic biases present in the ground truth or being introduced into the stress prediction model. Since the RF predictions were estimates of the unmodified ground truth stress scores, the convergence of the RF predictions to perceived stress is more accurately represented by the 0.19 (to 0.25 with state-trait anxiety) correlation with ground truth (see Table 3).

This raises the question of how to interpret the magnitude of the correlations. Using the widely used, but sometimes disputed, criteria of 0.1, 0.3, and 0.5 correlations corresponding to small, medium, and large effects [126], the present results are suggestive of a small to medium-sized effect. Put differently, our correlations are equivalent to Cohen's d's of 0.39 and 0.52, respectively, which are also within the small-to-medium sized range in terms of effect sizes. Though modest, these results are consistent with decades of research in the psychological sciences pertaining to weak associations between behavioral signals and subjective mental states, especially when the phenomena occur naturally rather than being experimentally elicited [127], [128]. Consider for example, a meta-analysis of 36 lab-studies that contrasted resting state HRV in 2086 patients with an anxiety disorder and 2294 healthy controls [129]. This study concluded that "anxiety disorders are associated with reduced HRV [and] associated with a small-to-moderate effect size" (Hedges' $g = -0.29$ for high frequency HRV, CI: [-0.41

to -0.17], $p < 0.001$), which is consistent with the effects found here.

In summary, the limited amount of information relevant to perceived stress present in the multitude of sensors and features we gathered in this work strongly suggests that there is no uncomplicated path to automated stress inference, which should not be surprising given the multitude of factors and stressors which influence each of us daily. Indeed, there is a major conceptual leap between lab studies with elicited affect, research-grade sensors, and clear ground truth and studies conducted in-the-wild with natural affect, consumer-grade sensors, and noisy ground truth. When developing new stress prediction techniques, especially for inferring stress scores directly rather than (over-)simplified versions of the problem, we should expect performance gains to be roughly proportional to our intuition and expectations about stress. Given the low to moderate correlations associated with the valid and robust approach to perceived stress prediction in this work, we want to emphasize that the problem of developing real-world daily stress inference systems based on passive sensing is still largely unsolved. We hope the results in this work can serve as a foundation for future efforts attempting to link physiological, behavioral, and contextual information to perceived stress in natural settings.

5.2 Limitations and Future Work

Though this work strived to illustrate how reliably and robustly daily perceived stress could be predicted using physiological, contextual, and environmental information, there were a number of limitations. This work predicted daily stress from aggregated measures over different periods of time during the day, but it was unable to model the temporal dynamics (e.g. motifs) of stress during a single day due to the low temporal resolution of stress labels. The importance of the "at work today" feature in Figure 5 emphasizes that stress is higher on work days, but more frequent stress labels would enable models to determine which factors conditionally contribute to stress when at work compared to other times. Several recent works underscore that the performance of perceived stress inference can be improved when more contextual information (e.g., activity labels, day of the week) is combined with higher temporal resolution perceived stress reports [34], [35]. There were many observable signals that were not captured or were captured with low fidelity in our data that may offer further insights into perceived stress dynamics. Some examples include vocalizations and speech, traffic reports, and work-specific stressors such as email or distractions. Furthermore, the recent COVID-19 pandemic has affected, perhaps permanently, the times and locations where people engage in their work, which merits further investigation.

To complement this endeavor, adequate time series modeling techniques and experimentation will be needed. The two time-aware models tested in this paper (GRU and LSTM) performed at chance compared to the MLP model which they were based on, likely due to the lack of sufficient time series data. In separate experiments not reported in this paper, we tried training the MLP, GRU, and LSTM models for 500 epochs instead of 50 to see if performance improved. None of these models benefited from the additional training, so more work will be needed to test these and other promising time-aware models that might leverage the rich temporal and contextual information gained from future data sets supporting stress inference in the wild.

One intriguing avenue for further research is developing an understanding of how common contextual factors (e.g. physical and mental activities, location) specifically influence stress. Some research is already striving towards this goal [34], [35], [38], but the results so far are limited to binary or binned stress prediction. More work is needed to understand how much a situational context differentially affects stress levels within individuals and how to use this information to accurately and robustly infer stress levels.

Further research is also needed to understand the relationship between the number of compliance days and perceived stress. The negative relationship we observed ($\beta = -0.14$ from Table S3) suggests the absence of data (i.e. non-compliance) provides valuable contextual information for stress prediction. Future work should consider modeling patterns in missing data and inferring when the lack of data can be interpreted in this way.

Arguably, the ultimate goal of achieving robust and valid stress inference in the wild is to provide feedback to participants about stress levels when that stress is having a negative effect on well-being and performance. Future work focused on this topic should seek to understand when and how much stress is beneficial versus detrimental. Finding the right time and right method for providing feedback to negatively stressed individuals will also be difficult because, among other factors, different people prefer feedback at different times and in different ways [130] and simply suggesting that people are stressed may induce additional stress [131], [132].

6 Concluding Remarks

Robust and valid prediction of perceived stress *in situ* and at scale using contextual information and data from wearables depends on the procedural faithfulness of the modeling pipeline used to produce a trained learning model. In our case study (N=606), we were able to achieve a 0.19 Spearman correlation (equivalent to a *Cohen's d* of 0.39), which is consistent with a small-to-medium effect size [133]. Our inclusion of personal difference information (i.e., state-trait anxiety) led to an effect size increase of 0.11, demonstrating that individual differences in sensitivity to these stressors likely plays a significant role. We believe that the results presented in this study advance the field of *in situ* stress assessment by offering a practical, robust, and valid benchmark for daily stress inference using an assortment of contextual and physiological information collected from a large diverse group.

References

[1] W. J. Kop, N. J. Weissman, J. Zhu, R. W. Bonsall, M. Doyle, M. R. Stretch, S. B. Glaes, D. S. Krantz, J. S. Gottdiener, and R. P. Tracy, "Effects of acute mental stress and exercise on inflammatory markers in patients with coronary artery disease and healthy controls," *The American journal of cardiology*, vol. 101, no. 6, pp. 767–773, 2008.

[2] C. J. Mulligan, "Early environments, stress, and the epigenetics of human health," *Annual Review of Anthropology*, vol. 45, pp. 233–249, 2016.

[3] I. Niedhammer, S. David, S. Degioanni, A. Drummond, P. Philip, and . O. Physicians, "Workplace bullying and sleep disturbances: findings from a large scale cross-sectional survey in the french working population," *Sleep*, vol. 32, no. 9, pp. 1211–1219, 2009.

[4] A. Knutsson, "Health disorders of shift workers," *Occupational medicine*, vol. 53, no. 2, pp. 103–108, 2003.

[5] P. Ferri, M. Guadi, L. Marcheselli, S. Balduzzi, D. Magnani, and R. Di Lorenzo, "The impact of shift work on the psychological and physical health of nurses in a general hospital: a comparison between rotating night shifts and day shifts," *Risk management and healthcare policy*, vol. 9, p. 203, 2016.

[6] Nov 2018. [Online]. Available: https://www.kornferry.com/insights/articles/workplace-stress-motivation

[7] J. A. Okely, A. Weiss, and C. R. Gale, "The interaction between stress and positive affect in predicting mortality," *Journal of psychosomatic research*, vol. 100, pp. 53–60, 2017.

[8] A. DeLongis, J. C. Coyne, G. Dakof, S. Folkman, and R. S. Lazarus, "Relationship of daily hassles, uplifts, and major life events to health status." *Health psychology*, vol. 1, no. 2, p. 119, 1982.

[9] L. Eiland and B. S. McEwen, "Early life stress followed by subsequent adult chronic stress potentiates anxiety and blunts hippocampal structural remodeling," *Hippocampus*, vol. 22, no. 1, pp. 82–91, 2012.

[10] S. J. Lupien, B. S. McEwen, M. R. Gunnar, and C. Heim, "Effects of stress throughout the lifespan on the brain, behaviour and cognition," *Nature reviews neuroscience*, vol. 10, no. 6, pp. 434–445, 2009.

[11] B. S. McEwen and E. Stellar, "Stress and the individual: mechanisms leading to disease," *Archives of internal medicine*, vol. 153, no. 18, pp. 2093–2101, 1993.

[12] H. M. Lagraauw, J. Kuiper, and I. Bot, "Acute and chronic psychological stress as risk factors for cardiovascular disease: Insights gained from epidemiological, clinical and experimental studies," *Brain, behavior, and immunity*, vol. 50, pp. 18–30, 2015.

[13] C. Hammen, "Stress and depression," *Annu. Rev. Clin. Psychol.*, vol. 1, pp. 293–319, 2005.

[14] M. Basta, G. P. Chrousos, A. Vela-Bueno, and A. N. Vgontzas, "Chronic insomnia and the stress system," *Sleep medicine clinics*, vol. 2, no. 2, pp. 279–291, 2007.

[15] I. Antonijevic and I. Antonijevic, "Hpa axis and sleep: identifying subtypes of major depression," *Stress*, vol. 11, no. 1, pp. 15–27, 2008.

[16] R. C. Kessler, "The effects of stressful life events on depression," *Annual review of psychology*, vol. 48, no. 1, pp. 191–214, 1997.

[17] N. Slopen, D. R. Williams, G. M. Fitzmaurice, and S. E. Gilman, "Sex, stressful life events, and adult onset depression and alcohol dependence: are men and women equally vulnerable?" *Social Science & Medicine*, vol. 73, no. 4, pp. 615–622, 2011.

[18] S. L. Sommerfeldt, S. M. Schaefer, M. Brauer, C. D. Ryff, and R. J. Davidson, "Individual differences in the association between subjective stress and heart rate are related to psychological and physical well-being," *Psychological science*, vol. 30, no. 7, pp. 1016–1029, 2019.

[19] L. Varvogli and C. Darviri, "Stress management techniques: evidence-based procedures that reduce stress and promote health," *Health science journal*, vol. 5, no. 2, p. 74, 2011.

[20] "Highlights: Workplace stress & anxiety disorders survey." [Online]. Available: https://adaa.org/workplace-stress-anxiety-disorders-survey

[21] S. Gedam and S. Paul, "Automatic stress detection using wearable sensors and machine learning: A review," in *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. IEEE, 2020, pp. 1–7.

[22] A. Shaw, N. Simsiri, I. Deznaby, M. Fiterau, and T. Rahaman, "Personalized student stress prediction with deep multitask network," *arXiv preprint arXiv:1906.11356*, 2019.

[23] B. Bent, B. A. Goldstein, W. A. Kibbe, and J. P. Dunn, "Investigating sources of inaccuracy in wearable optical heart rate sensors," *NPJ digital medicine*, vol. 3, no. 1, pp. 1–9, 2020.

[24] J. Kim, A. S. Campbell, B. E.-F. de Ávila, and J. Wang, "Wearable biosensors for healthcare monitoring," *Nature biotechnology*, vol. 37, no. 4, pp. 389–406, 2019.

[25] D. R. Seshadri, R. T. Li, J. E. Voos, J. R. Rowbottom, C. M. Alfes, C. A. Zorman, and C. K. Drummond, "Wearable sensors for monitoring the internal and external workload of the athlete," *NPJ digital medicine*, vol. 2, no. 1, pp. 1–18, 2019.

[26] Y. Wurmser, "Wearables 2019-emarketer trends, forecasts & statistics," *Retrieved September*, vol. 7, p. 2019, 2019.

[27] Y. S. Can, N. Chalabianloo, D. Ekiz, and C. Ersoy, "Continuous stress detection using wearable sensors in real life: Algorithmic programming contest case study," *Sensors*, vol. 19, no. 8, p. 1849, 2019.

[28] H. J. Baek and J. Shin, "Effect of missing inter-beat interval data on heart rate variability analysis using wrist-worn wearables," *Journal of Medical Systems*, vol. 41, no. 10, p. 147, 2017.

[29] J. Choi and R. Gutierrez-Osuna, "Using heart rate monitors to detect mental stress," in *2009 Sixth International Workshop on Wearable and Implantable Body Sensor Networks*. IEEE, 2009, pp. 219–223.

[30] Y. Lutchyn, P. Johns, M. Czerwinski, S. Iqbal, G. Mark, and A. Sano, "Stress is in the eye of the beholder," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2015, pp. 119–124.

[31] J. M. Peake, G. Kerr, and J. P. Sullivan, "A critical review of consumer wearables, mobile applications, and equipment for providing biofeedback, monitoring stress, and sleep in physically active populations," *Frontiers in physiology*, vol. 9, p. 743, 2018.

[32] V. Sandulescu, S. Andrews, D. Ellis, N. Bellotto, and O. M. Mozos, "Stress detection using wearable physiological sensors," in *International Work-Conference on the Interplay Between Natural and Artificial Computation*. Springer, 2015, pp. 526–532.

[33] A. Sano, A. J. Phillips, Z. Y. Amy, A. W. McHill, S. Taylor, N. Jaques, C. A. Czeisler, E. B. Klerman, and R. W. Picard, "Recognizing academic performance, sleep quality, stress level, and mental health using personality traits, wearable sensors and mobile phones," in *2015 IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*. IEEE, 2015, pp. 1–6.

[34] V. Mishra, T. Hao, S. Sun, K. N. Walter, M. J. Ball, C.-H. Chen, and X. Zhu, "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*, 2018, pp. 1708–1716.

[35] M. Gjoreski, H. Gjoreski, M. Luštrek, and M. Gams, "Continuous stress detection using a wrist device: in laboratory and real life," in *proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing: Adjunct*, 2016, pp. 1185–1193.

[36] E. Smets, E. Rios Velazquez, G. Schiavone, I. Chakroun, E. D'Hondt, W. De Raedt, J. Cornelis, O. Janssens, S. Van Hoecke, S. Claes *et al.*, "Large-scale wearable data reveal digital phenotypes for daily-life stress detection," *NPJ digital medicine*, vol. 1, no. 1, pp. 1–10, 2018.

[37] M. Soto, C. Satterfield, T. Fritz, G. C. Murphy, D. C. Shepherd, and N. Kraft, "Observing and predicting knowledge worker stress, focus and awakeness in the wild," *International Journal of Human-Computer Studies*, vol. 146, p. 102560, 2021.

[38] Y. S. Can, D. Gokay, D. R. Kılıç, D. Ekiz, N. Chalabianloo, and C. Ersoy, "How laboratory experiments can be exploited for monitoring stress in the wild: a bridge between laboratory and daily life," *Sensors*, vol. 20, no. 3, p. 838, 2020.

[39] W. Lovallo, "The cold pressor test and autonomic function: a review and integration," *Psychophysiology*, vol. 12, no. 3, pp. 268–282, 1975.

[40] Z. B. Moses, L. J. Luecken, and J. C. Eason, "Measuring task-related changes in heart rate variability," in *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2007, pp. 644–647.

[41] C. Kirschbaum, K.-M. Pirke, and D. H. Hellhammer, "The 'trier social stress test'–a tool for investigating psychobiological stress responses in a laboratory setting," *Neuropsychobiology*, vol. 28, no. 1-2, pp. 76–81, 1993.

[42] P. Jönsson, M. Wallergård, K. Österberg, Å. M. Hansen, G. Johansson, and B. Karlson, "Cardiovascular and cortisol reactivity and habituation to a virtual reality version of the trier social stress test: a pilot study," *Psychoneuroendocrinology*, vol. 35, no. 9, pp. 1397–1403, 2010.

[43] V. Engert, S. I. Efanov, A. Duchesne, S. Vogel, V. Corbo, and J. C. Pruessner, "Differentiating anticipatory from reactive cortisol responses to psychosocial stress," *Psychoneuroendocrinology*, vol. 38, no. 8, pp. 1328–1337, 2013.

[44] J. Campbell and U. Ehlert, "Acute psychosocial stress: does the emotional stress response correspond with physiological responses?" *Psychoneuroendocrinology*, vol. 37, no. 8, pp. 1111–1134, 2012.

[45] H. Selye, "A syndrome produced by diverse nocuous agents," *Nature*, vol. 138, no. 3479, pp. 32–32, 1936.

[46] N. Schneiderman, G. Ironson, and S. D. Siegel, "Stress and health: psychological, behavioral, and biological determinants," *Annu. Rev. Clin. Psychol.*, vol. 1, pp. 607–628, 2005.

[47] M. L. Harris, C. Oldmeadow, A. Hure, J. Luu, D. Loxton, and J. Attia, "Stress increases the risk of type 2 diabetes onset in women: A 12-year longitudinal study using causal modelling," *PloS one*, vol. 12, no. 2, p. e0172126, 2017.

[48] S. Folkman and R. S. Lazarus, *Stress, appraisal, and coping*. New York: Springer Publishing Company, 1984.

[49] H.-G. Kim, E.-J. Cheon, D.-S. Bai, Y. H. Lee, and B.-H. Koo, "Stress and heart rate variability: a meta-analysis and review of the literature," *Psychiatry investigation*, vol. 15, no. 3, p. 235, 2018.

[50] A. J. Oldehinkel, J. Ormel, N. M. Bosch, E. M. Bouma, A. M. Van Roon, J. G. Rosmalen, and H. Riese, "Stressed out? associations between perceived and physiological stress responses in adolescents: The trails study," *Psychophysiology*, vol. 48, no. 4, pp. 441–452, 2011.

[51] S. S. Walvekar, J. G. Ambekar, and B. B. Devaranavadagi, "Study on serum cortisol and perceived stress scale in the police constables," *Journal of clinical and diagnostic research: JCDR*, vol. 9, no. 2, p. BC10, 2015.

[52] M. Van Eck, H. Berkhof, N. Nicolson, and J. Sulon, "The effects of perceived stress, traits, mood states, and stressful daily events on salivary cortisol," *Psychosomatic medicine*, vol. 58, no. 5, pp. 447–458, 1996.

[53] R. S. Lazarus, *Stress and emotion: A new synthesis*. Springer publishing company, 2006.

[54] D. D. Colgan, D. Klee, T. Memmott, J. Proulx, and B. Oken, "Perceived stress mediates the relationship between mindfulness and negative affect variability: A randomized controlled trial among middle-aged to older adults," *Stress and Health*, vol. 35, no. 1, pp. 89–97, 2019.

[55] J. Hellhammer and M. Schubert, "The physiological response to trier social stress test relates to subjective measures of stress during but not before or after the test," *Psychoneuroendocrinology*, vol. 37, no. 1, pp. 119–124, 2012.

[56] A. Bali and A. S. Jaggi, "Clinical experimental stress studies: methods and assessment," *Reviews in the Neurosciences*, vol. 26, no. 5, pp. 555–579, 2015.

[57] G. Huether, S. Doering, U. Rüger, E. Rüther, and G. Schüssler, "The stress-reaction process and the adaptive modification and reorganization of neuronal networks," *Psychiatry Research*, vol. 87, no. 1, pp. 83–95, 1999.

[58] R. P. Sloan, P. Shapiro, E. Bagiella, J. Bigger, E. Lo, and J. Gorman, "Relationships between circulating catecholamines and low frequency heart period variability as indices of cardiac sympathetic activity during mental stress," *Psychosomatic Medicine*, vol. 58, no. 1, pp. 25–31, 1996.

[59] H. Ruediger, R. Seibt, K. Scheuch, M. Krause, and S. Alam, "Sympathetic and parasympathetic activation in heart rate variability in male hypertensive patients under mental stress," *Journal of human hypertension*, vol. 18, no. 5, pp. 307–315, 2004.

[60] Y. Liu and S. Du, "Psychological stress level detection based on electrodermal activity," *Behavioural brain research*, vol. 341, pp. 50–53, 2018.

[61] C. Schubert, M. Lambertz, R. Nelesen, W. Bardwell, J.-B. Choi, and J. Dimsdale, "Effects of stress on heart rate complexity—a comparison between short-term and chronic stress," *Biological psychology*, vol. 80, no. 3, pp. 325–332, 2009.

[62] T. M. Spruill, "Chronic psychosocial stress and hypertension," *Current hypertension reports*, vol. 12, no. 1, pp. 10–16, 2010.

[63] K. Yamanaka and M. Kawakami, "Convenient evaluation of mental stress with pupil diameter," *International journal of occupational safety and ergonomics*, vol. 15, no. 4, pp. 447–450, 2009.

[64] T. Mano, "Microneurography as a tool to investigate sympathetic nerve responses to environmental stress." *Aviakosmicheskaia i ekologicheskaia meditsina= Aerospace and environmental medicine*, vol. 31, no. 1, pp. 8–14, 1997.

[65] S. Cohen, T. Kamarck, and R. Mermelstein, "A global measure of perceived stress," *Journal of health and social behavior*, pp. 385–396, 1983.

[66] D. M. Campagne, "Stress and perceived social isolation (loneliness)," *Archives of gerontology and geriatrics*, vol. 82, pp. 192–199, 2019.

[67] K. Bhui, S. Dinos, M. Galant-Miecznikowska, B. de Jongh, and S. Stansfeld, "Perceptions of work stress causes and effective interventions in employees working in public, private and non-governmental organisations: a qualitative study," *BJPsych bulletin*, vol. 40, no. 6, pp. 318–325, 2016.

[68] S. A. Taylor, N. Jaques, E. Nosakhare, A. Sano, and R. Picard, "Personalized multitask learning for predicting tomorrow's mood, stress, and health," *IEEE Transactions on Affective Computing*, 2017.

[69] B. M. Booth, K. Mundnich, T. Feng, A. Nadarajan, T. H. Falk, J. L. Villatte, E. Ferrara, and S. Narayanan, "Multimodal human and environmental sensing for longitudinal behavioral studies in naturalistic settings: Framework for sensor selection, deployment, and management," *Journal of medical Internet research*, vol. 21, no. 8, p. e12832, 2019.

[70] E. Adam, *Spit, sweat, and tears: Measuring biological data in naturalistic settings*. University of Michigan Press, 2009, pp. 3–37.

[71] R. Wang, F. Chen, Z. Chen, T. Li, G. Harari, S. Tignor, X. Zhou, D. Ben-Zeev, and A. T. Campbell, "Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones," in *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing*, 2014, pp. 3–14.

[72] K. Mundnich, B. M. Booth, M. l'Hommedieu, T. Feng, B. Girault, J. L'hommedieu, M. Wildman, S. Skaaden, A. Nadarajan, J. L. Villatte *et al.*, "Tiles-2018, a longitudinal physiologic and behavioral data set of hospital workers," *Scientific Data*, vol. 7, no. 1, pp. 1–26, 2020.

[73] S. M. Mattingly, J. M. Gregg, P. Audia, A. E. Bayraktaroglu, A. T. Campbell, N. V. Chawla, V. Das Swain, M. De Choudhury, S. K. D'Mello, A. K. Dey *et al.*, "The tesserae project: Large-scale, longitudinal, in situ, multimodal sensing of information workers," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–8.

[74] N. L. Sin, R. P. Sloan, P. S. McKinley, and D. M. Almeida, "Linking daily stress processes and laboratory-based heart rate variability in a national sample of midlife and older adults," *Psychosomatic medicine*, vol. 78, no. 5, p. 573, 2016.

[75] E. Hynynen, N. Konttinen, U. Kinnunen, H. Kyröläinen, and H. Rusko, "The incidence of stress symptoms and heart rate variability during sleep and orthostatic test," *European journal of applied physiology*, vol. 111, no. 5, pp. 733–741, 2011.

[76] G. Vilagut, "Test-retest reliability," *Encyclopedia of quality of life and well-being research*, pp. 6622–6625, 2014.

[77] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *arXiv preprint arXiv:1908.09635*, 2019.

[78] A. Sano, S. Taylor, A. W. McHill, A. J. Phillips, L. K. Barger, E. Klerman, and R. Picard, "Identifying objective physiological markers and modifiable behaviors for self-reported stress and mental health status using wearable sensors and mobile phones: observational study," *Journal of medical Internet research*, vol. 20, no. 6, p. e210, 2018.

[79] T. Umematsu, A. Sano, S. Taylor, and R. W. Picard, "Improving students' daily life stress forecasting using lstm neural networks," in *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE, 2019, pp. 1–4.

[80] G. Mark, S. T. Iqbal, M. Czerwinski, P. Johns, A. Sano, and Y. Lutchyn, "Email duration, batching and self-interruption: Patterns of email use on productivity and stress," in *Proceedings of the 2016 CHI conference on human factors in computing systems*, 2016, pp. 1717–1728.

[81] A. Sano, P. Johns, and M. Czerwinski, "Designing opportune stress intervention delivery timing using multi-modal data," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2017, pp. 346–353.

[82] A. Anusha, P. Sukumaran, V. Sarveswaran, A. Shyam, T. J. Akl, S. Preejith, M. Sivaprakasam *et al.*, "Electrodermal activity based pre-surgery stress detection using a wrist wearable," *IEEE journal of biomedical and health informatics*, vol. 24, no. 1, pp. 92–100, 2019.

[83] Z. D. King, J. Moskowitz, B. Egilmez, S. Zhang, L. Zhang, M. Bass, J. Rogers, R. Ghaffari, L. Wakschlag, and N. Alshurafa, "Micro-stress ema: A passive sensing framework for detecting in-the-wild stress in pregnant mothers," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 3, pp. 1–22, 2019.

[84] M. Dietz, I. Aslan, D. Schiller, S. Flutura, A. Steinert, R. Klebbe, and E. André, "Stress annotations from older adults-exploring the foundations for mobile ml-based health assistance," in *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare*, 2019, pp. 149–158.

[85] A. Sano and R. W. Picard, "Stress recognition using wearable sensors and mobile phones," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, 2013, pp. 671–676.

[86] M. Gjoreski, H. Gjoreski, M. Lutrek, and M. Gams, "Automatic detection of perceived stress in campus students using smartphones," in *2015 International Conference on Intelligent Environments*. IEEE, 2015, pp. 132–135.

[87] A. Tiwari, S. Narayanan, and T. H. Falk, "Stress and anxiety measurement" in-the-wild" using quality-aware multi-scale hrv features," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019, pp. 7056–7059.

[88] A. S. Dissing, T. B. Jørgensen, T. A. Gerds, N. H. Rod, and R. Lund, "High perceived stress and social interaction behaviour among young adults. a study based on objective measures of face-to-face and smartphone interactions," *PloS one*, vol. 14, no. 7, p. e0218429, 2019.

[89] E. N. Smith, E. Santoro, N. Moraveji, M. Susi, and A. J. Crum, "Integrating wearables in stress management interventions: Promising evidence from a randomized trial." *International Journal of Stress Management*, vol. 27, no. 2, p. 172, 2020.

[90] C. D. Spielberger, *State-trait anxiety inventory: a comprehensive bibliography*. Consulting Psychologists Press, 1989.

[91] T. R. Kirchner and S. Shiffman, "Ecological momentary assessment." 2013.

[92] G. N. Yannakakis, R. Cowie, and C. Busso, "The ordinal nature of emotions: An emerging approach," *IEEE Transactions on Affective Computing*, 2018.

[93] J. Allen, "Photoplethysmography and its application in clinical physiological measurement," *Physiological measurement*, vol. 28, no. 3, p. R1, 2007.

[94] S. R. Pasadyn, M. Soudan, M. Gillinov, P. Houghtaling, D. Phelan, N. Gillinov, B. Bittel, and M. Y. Desai, "Accuracy of commercially available heart rate monitors in athletes: a prospective study," *Cardiovascular diagnosis and therapy*, vol. 9, no. 4, p. 379, 2019.

[95] K. Georgiou, A. V. Larentzakis, N. N. Khamis, G. I. Alsuhaibani, Y. A. Alaska, and E. J. Giallafos, "Can wearable devices accurately measure heart rate variability? a systematic review," *Folia medica*, vol. 60, no. 1, pp. 7–20, 2018.

[96] S. Stevens and C. Siengsukon, "Commercially-available wearable provides valid estimate of sleep stages (p3. 6-042)," 2019.

[97] P. Robles-Granda, S. Lin, X. Wu, S. D'Mello, G. J. Martinez, K. Saha, K. Nies, G. Mark, A. T. Campbell, M. De Choudhury *et al.*, "Jointly predicting job performance, personality, cognitive ability, affect, and well-being," *arXiv preprint arXiv:2006.08364*, 2020.

[98] S. H. Baik, R. S. Fox, S. D. Mills, S. C. Roesch, G. R. Sadler, E. A. Klonoff, and V. L. Malcarne, "Reliability and validity of the perceived stress scale-10 in hispanic americans with english or spanish language preference," *Journal of health psychology*, vol. 24, no. 5, pp. 628–639, 2019.

[99] S. Anwer, M. D. Manzar, A. H. Alghadir, M. Salahuddin, and U. A. Hameed, "Psychometric analysis of the perceived stress scale among healthy university students," *Neuropsychiatric Disease and Treatment*, vol. 16, p. 2389, 2020.

[100] E. M. Jackson, "Stress relief: The role of exercise in stress management," *ACSM's Health & Fitness Journal*, vol. 17, no. 3, pp. 14–19, 2013.

[101] L. E. Charles, J. E. Slaven, A. Mnatsakanova, C. Ma, J. M. Violanti, D. Fekedulegn, M. E. Andrew, B. J. Vila, and C. M. Burchfiel, "Association of perceived stress with sleep duration and sleep quality in police officers," *International journal of emergency mental health*, vol. 13, no. 4, p. 229, 2011.

[102] C. R. Rebello, P. B. Kallingappa, and P. G. Hegde, "Assessment of perceived stress and association with sleep quality and attributed stressors among 1st-year medical students: A cross-sectional study from karwar, karnataka, india," *Tzu-Chi Medical Journal*, vol. 30, no. 4, p. 221, 2018.

[103] S. Yoon, S.-s. Lee, J.-m. Lee, and K. Lee, "Understanding notification stress of smartphone messenger app," in *CHI'14 Extended Abstracts on Human Factors in Computing Systems*, 2014, pp. 1735–1740.

[104] V. Apaolaza, P. Hartmann, C. D'Souza, and A. Gilsanz, "Mindfulness, compulsive mobile social media use, and derived stress: The mediating roles of self-esteem and social anxiety," *Cyberpsychology, Behavior, and Social Networking*, vol. 22, no. 6, pp. 388–396, 2019.

[105] M. Pielot and L. Rello, "Productive, anxious, lonely: 24 hours without push notifications," in *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*, 2017, pp. 1–11.

[106] K. Wong, A. H. Chan, and S. Ngan, "The effect of long working hours and overtime on occupational health: a meta-analysis of evidence from 1998 to 2018," *International journal of environmental research and public health*, vol. 16, no. 12, p. 2102, 2019.

[107] L. Dusselier, B. Dunn, Y. Wang, M. C. Shelley iI, and D. F. Whalen, "Personal, health, academic, and environmental predictors of stress for residence hall students," *Journal of American college health*, vol. 54, no. 1, pp. 15–24, 2005.

[108] L. Thorn, P. Evans, A. Cannon, F. Hucklebridge, and A. Clow, "Seasonal differences in the diurnal pattern of cortisol secretion in healthy participants and those with self-assessed seasonal affective disorder," *Psychoneuroendocrinology*, vol. 36, no. 6, pp. 816–823, 2011.

[109] D. I. Rifkin, M. W. Long, and M. J. Perry, "Climate change and sleep: A systematic review of the literature and conceptual framework," *Sleep medicine reviews*, vol. 42, pp. 3–9, 2018.

[110] N. Obradovich, R. Migliorini, S. C. Mednick, and J. H. Fowler, "Nighttime temperature and human sleep loss in a changing climate," *Science advances*, vol. 3, no. 5, p. e1601555, 2017.

[111] M. Hashizaki, H. Nakajima, T. Shiga, M. Tsutsumi, and K. Kume, "A longitudinal large-scale objective sleep data analysis revealed a seasonal sleep variation in the japanese population," *Chronobiology international*, vol. 35, no. 7, pp. 933–945, 2018.

[112] S. M. Mattingly, T. Grover, G. J. Martinez, T. Aledavood, P. Robles-Granda, K. Nies, A. Striegel, and G. Mark, "The effects of seasons and weather on sleep patterns measured through longitudinal multimodal sensing," *NPJ digital medicine*, vol. 4, no. 1, pp. 1–15, 2021.

[113] L. Gondara and K. Wang, "Mida: Multiple imputation using denoising autoencoders," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2018, pp. 260–272.

[114] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," *International journal of forecasting*, vol. 22, no. 4, pp. 679–688, 2006.

[115] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[116] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org.

[117] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[118] M. Malik, "Heart rate variability: Standards of measurement, physiological interpretation, and clinical use: Task force of the european society of cardiology and the north american society for pacing and electrophysiology," *Annals of Noninvasive Electrocardiology*, vol. 1, no. 2, pp. 151–181, 1996.

[119] L. J. Iny, J. Pecknold, B. E. Suranyi-Cadotte, B. Bernier, L. Luthe, N. Nair, and M. J. Meaney, "Studies of a neurochemical link between depression, anxiety, and stress from [3h] imipramine and [3h] paroxetine binding on human platelets," *Biological psychiatry*, vol. 36, no. 5, pp. 251–291, 1994.

[120] Z. Ahmed and S. H. Julius, "The relationship between depression, anxiety and stress among women college students." *Indian Journal of Health & Wellbeing*, vol. 6, no. 12, 2015.

[121] N. Daviu, M. R. Bruchas, B. Moghaddam, C. Sandi, and A. Beyeler, "Neurobiological links between stress and anxiety," *Neurobiology of stress*, vol. 11, p. 100191, 2019.

[122] G. Konstantopoulou, T. Iliou, K. Karaivazoglou, G. Iconomou, K. Assimakopoulos, and P. Alexopoulos, "Associations between (sub) clinical stress-and anxiety symptoms in mentally healthy individuals and in major depression: a cross-sectional clinical study," *BMC psychiatry*, vol. 20, no. 1, pp. 1–8, 2020.

[123] C. D. Spielberger, "State-trait anxiety inventory," *The Corsini encyclopedia of psychology*, pp. 1–1, 2010.

[124] A. Benetos, A. Rudnichi, F. Thomas, M. Safar, and L. Guize, "Influence of heart rate on mortality in a french population: role of age, gender, and blood pressure," *Hypertension*, vol. 33, no. 1, pp. 44–52, 1999.

[125] A. H. Zohar, C. R. Cloninger, R. McCraty *et al.*, "Personality and heart rate variability: exploring pathways from personality to cardiac coherence and health," *Open Journal of Social Sciences*, vol. 1, no. 06, p. 32, 2013.
[126] J. Cohen, "A power primer," *Psychological bulletin*, vol. 112, no. 1, p. 155, 1992.
[127] S. D'Mello, A. Kappas, and J. Gratch, "The affective computing approach to affect measurement," *Emotion Review*, vol. 10, no. 2, pp. 174–183, 2018.
[128] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, "Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements," *Psychological science in the public interest*, vol. 20, no. 1, pp. 1–68, 2019.
[129] J. A. Chalmers, D. S. Quintana, M. J. Abbott, A. H. Kemp *et al.*, "Anxiety disorders are associated with reduced heart rate variability: a meta-analysis," *Frontiers in psychiatry*, vol. 5, p. 80, 2014.
[130] C. M. Kelley and A. C. McLaughlin, "Individual differences in the benefits of feedback for learning," *Human Factors*, vol. 54, no. 1, pp. 26–35, 2012.
[131] B. Yu, M. Funk, J. Hu, Q. Wang, and L. Feijs, "Biofeedback for everyday stress management: A systematic review," *Frontiers in ICT*, vol. 5, p. 23, 2018.
[132] N. Moraveji, A. Adiseshan, and T. Hagiwara, "Breathtray: augmenting respiration self-regulation without cognitive deficit," in *CHI'12 Extended Abstracts on Human Factors in Computing Systems*, 2012, pp. 2405–2410.
[133] J. Cohen, *Statistical power analysis for the behavioral sciences*. Academic press, 2013.

Brandon M. Booth is a postdoctoral research associate at CU Boulder working in the Emotive Computing Lab. His research focuses on using multi-modal machine learning techniques to model human perception, behavior, and experiences and developing algorithms to reduce the impact of inadvertent human biases and errors. He has a diverse industry background researching, publishing, and developing video games, serious games, robots, computer vision and human-computer interactions systems, and geo-spatiotemporal visualizers.

Hana Vrzakova received her PhD in eye tracking, machine learning, and HCI at University of Eastern Finland (Cor Baayen Award 2020, Honorable Mentioned). In her postdoctoral research at CU Boulder (Emotive Computing Lab), she investigated physiological and behavioral synchrony underlying remote collaborative problem solving and negotiation. She served as Head of Medical R&D in Microsurgery Training Center, Kuopio University Hospital, and Research Director in the EMC Program, University of Eastern Finland.

Stephen M. Mattingly graduated from the University of Notre Dame with a doctorate in Cognition, Brain, and Behavior. He is interested in sleep and stress and how to measure these constructs better in daily life using wearables and other passive sensors. He is also interested in using passive sensors to improve sleep duration and quality and to reduce stress.

Gonzalo J. Martinez is a PhD candidate in the Department of Computer Science and Engineering at the University of Notre Dame. His research lies at the intersection of data science, human-computer interaction and behavioral studies. Using data from ubiquitous sensors like wearables, smartphones, beacons and social media, he combines machine learning with behavioral psychology to improve health and well-being. His current work uses sensors and social media data to understand the behaviors that support good workplace performance.

Louis Faust received his PhD in Computer Science at the University of Notre Dame. There, his research focused on machine learning applications for health and wellness, leveraging data streams generated through wearable technology to better understand health behaviors and outcomes. He is currently a Data Scientist at the Mayo Clinic.

Sidney K. D'Mello is an Associate Professor with a joint appointment in the Institute of Cognitive Science and Department of Computer Science at University of Colorado Boulder. His research focuses on applying multimodal machine learning to investigate the interplay between the cognitive and affective states of individuals and teams engaged in real-world activities. D'Mello has co-edited seven books and published almost 300 journal papers, book chapters, and conferences proceedings. He is an associate editor for Discourse Processes and PloS ONE.