This article has been accepted for publication in IEEE Internet of Things Journal. 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/JIOT.2024.3375299

SOSW: Stress Sensing with Off-the-shelf Smartwatches in the Wild

Kobiljon Toshnazarov, Uichin Lee, Byung Hyung Kim, Varun Mishra, Lismer Andres Caceres Najarro, Youngtae Noh

***Abstract*—Recent advances in wearable technology have led to the development of various methods for stress sensing in both controlled laboratory and real-life environments. However, existing methods often rely on specialized or expensive sensors that may not be easily accessible to the general population. In this study, we investigate the feasibility of using off-the-shelf smartwatches for stress detection in real-life scenarios. To achieve this, we propose *SOSW*, a comprehensive methodology for robust sensor data processing by considering both physiological and contextual data. *SOSW* employs a two-layer machine learning (ML) architecture. The first-layer ML model is trained and validated using carefully collected data under controlled laboratory conditions. The second-layer ML model is trained and validated using data collected in real-life settings. We conducted evaluations with 26 and 18 participants in controlled laboratory and real-life conditions, respectively. The results indicate that our methodology can successfully detect stressful events with an F-1 score of up to 0.84 in laboratory conditions and 0.71 in real-life scenarios using off-the-shelf smartwatches. The results are comparable to those achieved by the state of the art methods that rely on dedicated wearables.**

***Index Terms*—*SOSW*, stress, smartwatch, commodity, in the wild, field, context**

I. Introduction

Stress is the response of our body to an internal or external threat to its homeostasis [1]. It represents a defense mechanism that the body employs to maintain its internal stability, commonly referred to as the fight-or-flight response [2]. During the stress response, the body undergoes internal changes, including an increase in heart rate (HR), blood pressure, respiration, as well as improved oxygenation and nutrition to the brain, heart, and skeletal muscles, among other effects [3]. As a result, the senses become sharper, and attention increases, enhancing the organism's ability to cope more effectively with the stressful situation [4]. Among various physiological signals, the heart activity, which controls the flow of blood in the veins carrying oxygen and nutrients to various body parts, is a key indicator of the body's physiological stress response [5].

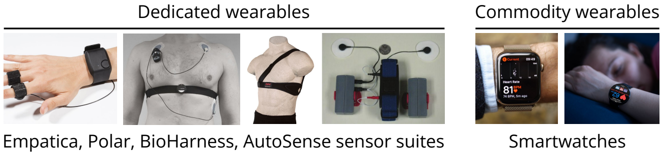

Fig. 1: Sensors commonly used for stress detection.

In daily life, our body utilizes the acute stress response, which involves short-lived changes in our physiology, in order to adapt to various situations [1], [6]. The acute stress response is used to tackle everyday challenges, enhancing performance, cognition, and memory in response to challenges and threats [1]. However, prolonged exposure to stressors and inadequate stress management can lead to malfunction in the stress response system, resulting in episodic acute and chronic stress [1]. Such sustained exposure to stressors can have a cumulative toll and has been associated with various health complications. For instance, it has been reported that chronic stress can cause cardiovascular problems [1], compromise the immune system [3], decrease work performance [7], and overall decrease the quality of life [3], [8]. Therefore, our work focuses on detecting acute stress in daily life settings to enable timely interventions, before developing further health complications.

A recent study revealed that stressful events that occur in our daily lives may lead to heterogeneous physiological responses [9] that are different from those observed in laboratory settings. This aligns with the perspective of emotion studies, which suggests that individuals' emotions may vary depending on various contexts (e.g., locations, social settings, and activities) [10]. Traditional studies leveraged controlled, laboratory-based settings to build a stress model based on the physiological responses. However, researchers warned that the practicability of such a model is limited [9], [11]. While laboratory-based models provide some hints for stress detection in the wild, it is important to capture physiological

This research was supported in part by the National Research Foundation (NRF) of Korea under grant number NRF-2021M3A9E4080780 and NRF-2021R111A1A01041257, Korean government (MOTIE) through the GAMS program under grant number G032653912 and through CEMS program under grant number G03002363251. This work was also supported by the KENTECH Research Grant (202200037A).

K. Toshnazarov (qobiljon@kentech.ac.kr) and L. A. Caceres Najarro (andrescn@kentech.ac.kr) are affiliated with Energy AI, KENTECH, South Korea.

U. Lee (uclee@kaist.ac.kr) is affiliated with the Korea Advanced Institute of Science & Technology (KAIST), South Korea.

B. H. Kim (bhyung@inha.ac.kr) is affiliated with Department of Artificial Intelligence, Inha University, South Korea.

V. Mishra (v.mishra@northeastern.edu) is affiliated with Khoury College of Computer Sciences, Northeastern University, United States.

Y. Noh (youngtaenoh@hanyang.ac.kr) is affiliated with Hanyang University, South Korea.

L. A. Caceres Najarro and Y. Noh are the corresponding authors.

responses under diverse stress episodes in the wild, and thus, prior studies used experience sampling methods where users are asked to self-report their perceived stress levels throughout the day [9], [11].

One key challenge is reliable data collection of physiological data in the wild [9], [11], [12], [13], [14], [15], [16]. Most of the prior studies relied on dedicated wearable HR sensors (e.g., Polar H7 and Bioharness) as shown in Fig. 1, which are bulky and inconvenient to wear [9], [17], [18]. These devices offer quality heart activity monitoring, but discomfort of wearing on a chest limits temporal coverage as well as adoption of a dedicated wearable hinders a large-scale data collection. This work aims to leverage general-purpose wearables like commodity smartwatches which offer passive, continuous HR and activity tracking as in Samsung Watch and Apple Watch, which further enable access to broader contextual information of users, such as activity detection, device use, social interactions, and health data. For this reason, we expect that smartwatches can achieve higher temporal coverage and broader contextual information in daily life scenarios.

Our objective is to prove that we can build stress model in real-life settings with commodity smartwatches in a reliable manner. This will help to realize real-time intervention apps using commodity smartwatches, helping people to better manage their stress in everyday contexts. For HR sensing, a common method is photoplethysmography (PPG) sensing which uses different wavelength lights and their reflection. However, PPG is known to be error-prone under motion (or physical activity) [19], [20] due to sensor displacement relative to skin and loose wearing conditions [18], [21]. It is important to understand the PPG sensing accuracy of a commodity smartwatch and to systematically study how to deal with PPG errors for model building. Therefore, our first research question is *RQ1: Is it possible for a commodity smartwatch to accurately identify physiological stress?*

Moreover, the population view of emotion argues that user's emotion depends on their contexts (e.g., places, social setting, and activities) [10], and it is very important to consider everyday contexts beyond the laboratory setting [22]. Beyond user's current physiological responses, we consider fusing multiple contextual data such as user's activities, location, social settings, device usage, mobility, among others. These additional data streams act as contextual signatures, which may help to improve the accuracy of stress model performance. We employ multi-sensor contextual data fusion to address our second research question *RQ2: What is the attainable level of accuracy in detecting stress by taking into account contextual data in the wild?*

While addressing these questions, we make the following four-fold contributions:

- We systematically analyze and demonstrate the limitations of off-the-shelf smartwatches for measuring individuals' HR in realistic scenarios.
- We develop a robust data processing pipeline that rigorously addresses limitations of smartwatch PPG sensing and incorporates diverse contextual data for real-life stress detection.
- This work combines physiological data collected from commercially available smartwatches with contextual information obtained from smartphones to accurately identify and measure stress levels in the wild.
- The extensive physiological and contextual dataset[1], together with the codebase[2] for our whole data processing pipeline, is accessible to the public for the rapid advancement of the scientific community.

We believe that *SOSW* makes a significant step towards stress detection in real-life settings using general-purpose wearables like smartwatches. Our methods enable smartwatches to achieve reliable physiological stress detection in laboratory settings, and their access to contextual data holds promise in significantly enhancing stress detection in the wild, surpassing the capabilities of dedicated wearables.

II. Background and related works

With recent advancements in sensor technology and mobile devices, it has become possible to collect physiological data and conduct experience sampling in real-life, free-living conditions using various devices, including smartphones, smartwatches, and smartbands. Previous efforts have employed both contact-based and contactless sensing methods, utilizing dedicated sensing wearables, to monitor physical and physiological signals of stress [23]. While dedicated sensing wearables (e.g., high end, custom-made, clinical-grade devices) offer high-quality sensor data in controlled environments like laboratory stress studies [24], [25], they also present practical challenges when used for daily-life stress tracking. These challenges include high costs, discomfort due to bulkiness, and limited market accessibility [9], [17], [18]. As a result, the use of such devices in real-life studies may raise questions about the reproducibility of findings [26]. An attempt to address this issue has been made by using commodity chest straps for stress tracking [9]. However, chest straps are invasive wearables that are not designed for daily use and may be unwelcomed by subjects, limiting its continuous sensing capability. In contrast, commodity smartwatches offer a practical solution due to their compact and non-invasive design, as illustrated in Fig. 1, leading to higher temporal coverage of physiological sensing in real-life scenarios. Therefore, we prioritize the use of commodity smartwatches for stress sensing in real-life settings.

When it comes to sensing physiological signals related to stress, previous research has demonstrated the feasibility of accurate stress detection using various physiological sensors, including PPG [13], [14], [27], [28], electrocardiography (ECG) [9], [11], [13], [29], galvanic skin response (GSR) [9], [13], [14], respiratory inductance plethysmogram (RIP)

[1] *SOSW* dataset: https://www.kaggle.com/datasets/kobiljon/sosw-ieee-iot
[2] *SOSW* codebase: https://github.com/qobiljon/sosw-pipeline

[11], [29], and skin temperature (ST) [14]. However, it is important to note that physiological sensors (e.g., PPG) have limitations in detecting perceived stress, as perceived stress may not always manifest as immediate physiological responses. Contextual information is known to significantly influence users' perceived stress levels, and when combined with physiological data, it can enhance stress detection accuracies in real-life scenarios [11], [22].

To enhance stress detection accuracy, recent efforts have explored the effectiveness of stress models initially trained in controlled laboratory settings when applied to field datasets containing real-life scenarios. This involves mapping pairs of physiological and perceived stress inferences from preceding data windows, typically the previous minute, using techniques like Bayesian networks or hidden Markov models [9], [11], [29]. However, it is essential to note that such approaches heavily rely on physiological stress response data obtained in controlled environments with limited types of stressors. These may not always be a perfect match for real-life scenarios, as different stressors can lead to diverse physiological responses [9]. For example, the mental arithmetic or socio-evaluative tasks in a laboratory protocol may not be directly applicable to situations like driving or other high-stress 'fight-or-flight' scenarios encountered in the wild.

Outside laboratory settings, there are few research efforts that employ commercially available wearables to detect stress in more naturalistic environments [30], [31]. For instance, the study [30] effectively applied physiological data of commercially available wrist wearables for stress detection of students during a 50-70 minute lecture. While the study offered valuable insights into the physiological stress experienced by students, it focused solely on a specific real-life situation, *i.e.,* a short-lived academic setting. Recently, another study [31] considered stress detection in uncontrolled daily life settings by leveraging commodity smartwatches paired with smartphones for the collection of physiological and contextual data, respectively. Although the study demonstrates a detection accuracy of 60 %, the authors overlooked the intrinsic limitations of optical-based PPG sensors found in smartwatches, which inevitably impact the accuracy of stress detection. Importantly, their field study sampled users' perceived stress labels at three specific scheduled times: 9am, 4pm, and 9pm. However, this approach may not fully capture the dynamics of perceived daily life stress, which can vary at random times throughout the day in natural conditions.

As opposed to these works, *SOSW* comprehensively addresses the limitations of smartwatch-based physiological signals and incorporates rich situational context data including previous contexts to detect perceived stress. Additionally, SOSW is not limited to a singular real-life scenario; rather, it encompasses a comprehensive approach to stress detection within a variety of naturalistic settings, effectively functioning in the wild. Furthermore, *SOSW* utilizes a physiological stress model trained in a controlled laboratory setting to estimate stress likelihood in real-life scenarios, a process we refer to as *laboratory knowledge transfer* to field. *SOSW* then incorporates this information, derived from the laboratory model's output, as an additional factor when training a model on the field dataset to provide insights into the user's physiological stress state, which further enhances the performance of *SOSW*.

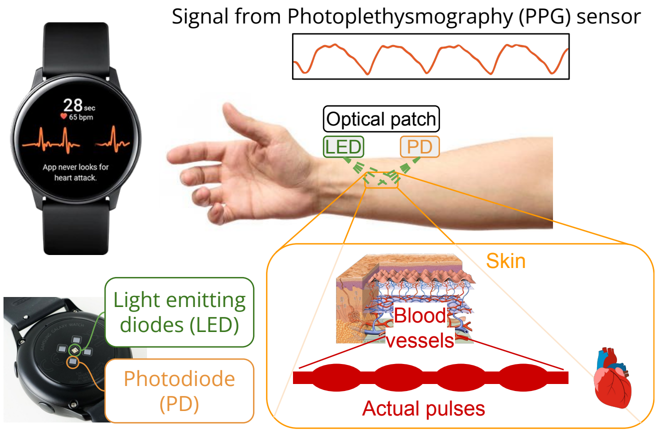
Signal from Photoplethysmography (PPG) sensor
28 sec
65 bpm
App never looks for heart attack.
Optical patch
LED
PD
Skin
Blood vessels
Light emitting diodes (LED)
Photodiode (PD)
Actual pulses

Fig. 2: Blood volume pulse signal acquisition by PPG sensor of commodity smartwatch.

Let us in passing add that although there exist commercial smartwatches that readily provide stress detection capabilities, they are limited in several aspects. For instance, Fitbit's stress management application presents summary of physiological readings to the user, and requires them to manually log their perceived stress levels instead of detecting it passively [32]. Also, Samsung smartwatches provide proprietary software that estimates physiological stress levels based on heart rate readings [33]. However, there are limited documentations on the proprietary software. Similarly, Garmin smartwatches also report stress levels derived from heart rate arousals [34]. Common limitation with these software is the limited accuracy of physiological stress detection and lack of perceived stress detection in real-life scenarios. Moreover, notably, even the widely popular Apple Watch does not yet offer stress detection features [35], [36]. In contrast to these limitations, our work aims to develop a comprehensive methodology for perceived stress detection in real-life scenarios. Rigorous evaluation of the methods involved in *SOSW* prove the efficacy of our methodology.

Since our study employs commodity smartwatches for acquiring physiological readings (*i.e.,* HR data) in real-life scenarios, it is highly important to systematically investigate the reliability of such data in realistic conditions. Like chest-worn sensors such as ECG and respiratory, it has been demonstrated that wrist-worn PPG sensors are also vulnerable to motion artifacts [19], [20], [37], [38], [39]. Prior works commonly rely on acceleration signals [40], [41], and various other PPG signal filtering methodologies [42], [43] to handle such data. Moreover, a recent study [21] argued that it is not enough to rely solely on motion sensors to handle PPG inaccuracies, reporting on how external lighting

This article has been accepted for publication in IEEE Internet of Things Journal. 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/JIOT.2024.3375299

TABLE I: Preliminary study settings, averaged across 28 participants. 'Motion' refers to the degree of motion during three different activities with natural body movements. 'Looseness' levels pertain to the tightness of the smartwatch strap based on the percentage ratio between wrist and strap circumferences.

Motion	**Static**	**Walking**	**Running**
	0 km/h	4 km/h	8 km/h
Looseness	**Tight**	**Medium**	**Loose**
	95.78% (±3.09)	92.12% (±3.19)	88.74% (±3.26)

conditions between skin and sensor can significantly affect PPG sensor accuracy as well. They quantified various sensor to wrist distances using artificial rings with different heights (*i.e.,* 3mm, 5mm, 7mm, and 5mm with holes). However, our work systematically revisits the importance of addressing both motion and looseness artifacts in more realistic mobility and wearing conditions.

III. Commodity smartwatch PPG inaccuracies – a preliminary study

A. Background and motivation

Commodity smartwatches utilize a PPG sensor that includes a light-emitting diode (LED) and photodiodes to capture physiological signals, such as HR. The LED emits light onto the skin, and the photodiodes measure the changes in the amount of reflected light, providing information about heart activity. In this process, as depicted in Fig. 2, the changes in the reflected light correspond to blood volume fluctuations in the wrist caused by the heartbeat. By analyzing the generated waveform, higher-level information from the users' heart activity such as the HR and interbeat interval (IBI) can be obtained. Such information, in particular the HR, is commonly used as input parameters for stress detection, thus accurate measurements of those parameters are of vital importance [44], [45]. Unfortunately, external light [18] and sensor displacement over the skin [19] can influence the accuracy of the PPG readings affecting the generated waveform and ultimately disrupting the accurate measurements of the users' heart activity.

Several studies have shown the susceptibility of PPG sensor measurements to motion and looseness artifacts [18], [19], [20], [21], [37], [38]. To address these challenges, existing works commonly rely on acceleration signals and various other PPG signal filtering methodologies [40], [41], [42]. The effectiveness of smartwatches in different wrist movements and when worn with a loose fit has been shown to have a significant negative impact on the optical measurements collected from these devices. Actions like as gripping, flexing, and extending fingers and wrist motions have been found to be particularly damaging in this regard [20], [46]. The impact of different sensor-to-skin distances and external light reaching the optical sensor between the smartwatch and the skin has been also analyzed. This analysis revealed that the PPG sensor light intensity variance increases when the smartwatch is worn loosely [18]. However, it is important to note that these studies did not consider both motion and looseness simultaneously in their experimental setups. Furthermore, some experiments involved the use of external objects, such as 3D-printed rings, to quantify the distance between the wrist skin and the smartwatch [21]. Although such a setup facilitated the evaluation of the impact of sensor-to-skin distance on HR measurements, the experiment does not resemble real-life scenarios.

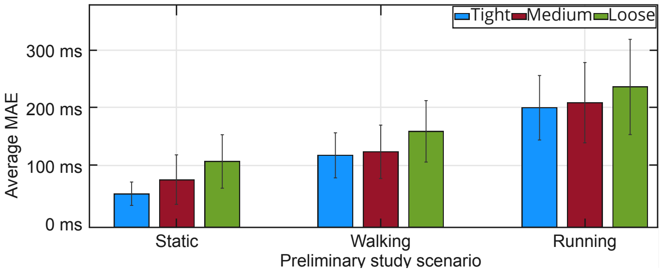

Fig. 3: Average mean absolute error (MAE) of IBI estimation by commodity smartwatch PPG sensor data across participants and scenarios. The x-axis represents the 9 different scenarios, and the y-axis represents the average MAE in IBI data estimation (*ms*) for the PPG sensor with reference to ECG-based data.

B. Preliminary study setup

The aim of our preliminary study is to explore two major limitations of commodity smartwatch PPG sensing: the impact of loose wearing conditions, and increased wrist motion intensities on the precision of HR estimation. This study involved a systematic analysis that replicated real-world scenarios by varying smartwatch strap looseness levels and wrist motion intensities, covering different realistic situations. The specific methods used for quantifying strap looseness levels and motion intensities, followed by our preliminary study scenarios are detailed below.

In our study, we categorized the levels of looseness of the smartwatch strap into three levels: tight, medium, and loose, based on the individual fit of the strap on each participant's wrist. To ensure accurate categorization, we initially measured each participant's wrist circumference, specifically at the point where the smartwatch is worn. The default strap that came with the smartwatch was utilized, with its tightness being adjusted according to each participant's wrist circumference. For precise classification, we identified the nearest strap hole that would bring the strap and wrist circumferences closest to each other, marking it as 'tight.' The subsequent two holes were marked as 'medium' and 'loose,' respectively. Our criteria for consistent looseness levels were established based on the ratio between the participant's wrist circumference and the smartwatch strap's circumference, represented as a percentage, as summarized in Table I.

This article has been accepted for publication in IEEE Internet of Things Journal. 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/JIOT.2024.3375299

With regards to quantifying motion intensity of wrist, we observed participants during three natural body movements, which took place on a standard chair and treadmill: static (*i.e.,* sitting on a chair), walking at a speed of 4 km/h, and running at 8 km/h, same as in a prior work [19]. To reduce potential sources of bias or confounding effects, we instructed participants to remain stationary while sitting on a chair, minimizing their wrist motion, and to maintain their natural wrist and body movements when using the treadmill, avoiding the use of handrails.

In summary, we combined three strap-looseness levels and three motion intensities, forming the following 9 scenarios: tight-static, tight-walking, tight-running, medium-static, medium-walking, medium-running, loose-static, loose-walking, and loose-running. Throughout these nine scenarios, we collected PPG-based HR data using the Samsung Watch 5 smartwatch [47] to evaluate its accuracy. Simultaneously, we obtained ECG-based HR data from the reliable Polar H10 chest strap [48], [49], which served as our reference or ground truth HR data. Each scenario, a 10-minute session, aimed to explore the accuracy of HR data of PPG sensor (with reference to ECG-based data) under a specific wrist motion intensity and smartwatch fit.

To explore the HR accuracy of commodity smartwatch, we recruited 28 participants, each of whom participated in the nine scenarios spanning approximately 90 minutes. To ensure that the results are not impacted by any external variables of the surrounding environment, our preliminary study was conducted indoors with consistent lighting, temperature, and humidity. We also shuffled the sequence of 9 scenarios and randomly assigned the smartwatch to the left or right wrist in order to prevent potential bias or confounding effects related to a specific scenario order or wrist placement, thus minimizing the potential for systematic errors.

C. *Preliminary study results*

The preliminary study results, summarized in Fig. 3, provide the average mean absolute error (MAE) of the smartwatch PPG sensor in estimating IBI data in milliseconds (*ms*) across the 28 participants.

The accuracy of HR estimation tended to decrease with higher levels of motion intensity. In contrast, the most precise IBI estimations were obtained in static conditions, *i.e.,* when wrist motion was minimal.

Interestingly, the ‘tight’ level of wearing the smartwatch strap did not mitigate the errors in IBI estimation accuracy caused by increased wrist motion intensities. This indicates that motion intensity remains the primary factor influencing IBI estimation accuracy. Furthermore, the impact of looseness levels on IBI data accuracy is most pronounced when participants are in a static condition, with no significant wrist movements. Moreover, a significant increase in IBI estimation error is observed as motion intensity rises, even when the smartwatch is securely fastened to the participants’ wrists.

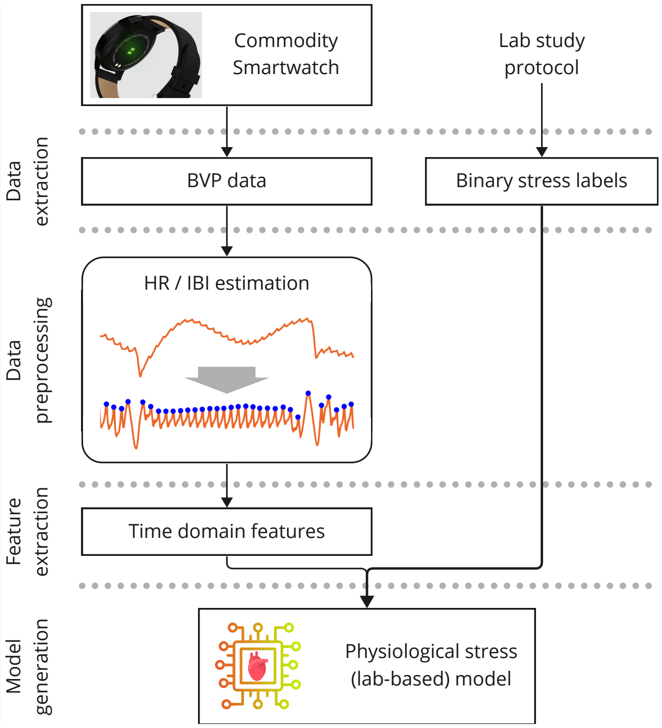

Fig. 4: Laboratory data processing pipeline involving PPG signals of commodity smartwatch.

Based on obtained results, relying on the smartwatch’s PPG-based HR estimation is most suitable when the participant remains stationary. Notably, even in a static condition, there is necessity to further process the PPG readings from the smartwatch for reliable use, addressing the looseness levels. As a result, we incorporate combination of motion and looseness filters in our methodology to significantly enhance the quality of the PPG readings. In the following section, we will delve into the specifics of these filters in the methodology.

IV. Methodology: *SOSW* Data Processing Pipeline

The ultimate objective of our methodology is to accurately detect stress in real-world settings using physiological data from smartwatches and contextual data from smartphones. To this end, we introduce *SOSW*, a robust and comprehensive data processing pipeline for laboratory and in-the-wild stress detection. This pipeline not only overcomes the limitations inherent in data from commodity smartwatches in real-life scenarios, but also enhances accuracy by integrating rich contextual data from situational environments.

Our methodology includes rigorous methods for handling commodity smartwatch PPG signals in both laboratory and real-life settings. This includes addressing challenges like motion and looseness artifacts in real-life scenarios. We train and validate a physiological stress detection model based on laboratory data, selecting the most accurate and reliable one. We utilize this model to gain insights into users’

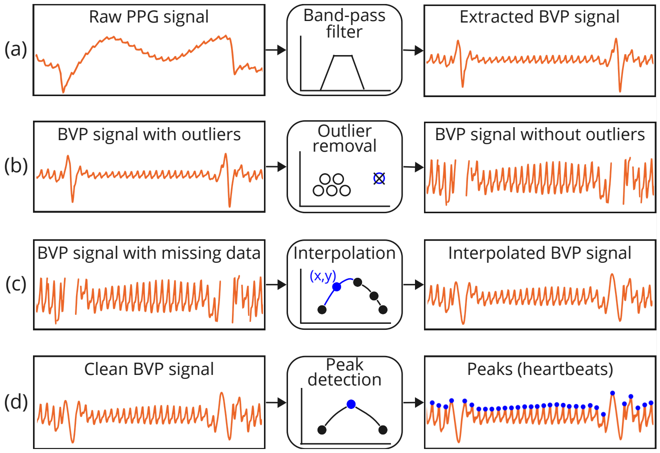

Fig. 5: Four steps of preprocessing of BVP data for physiological stress model development: (a) Band-pass filter, (b) Outlier removal, (c) Interpolation, and (d) Peak detection for HR estimation

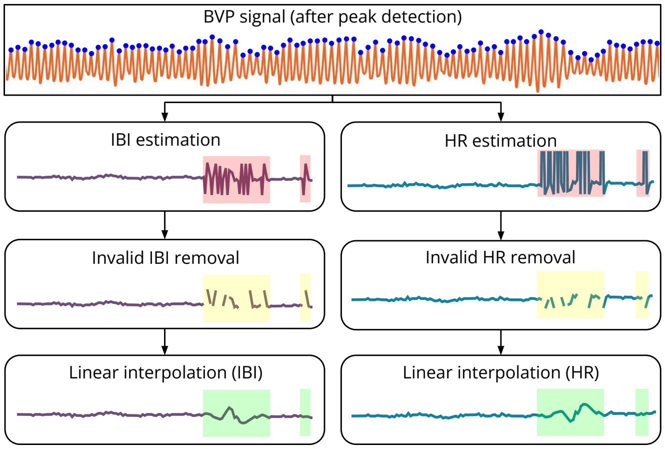
BVP signal (after peak detection)
IBI estimation
HR estimation
Invalid IBI removal
Invalid HR removal
Linear interpolation (IBI)
Linear interpolation (HR)

Fig. 6: Invalid IBI and HR removal. Red shaded area depicts the segment of data with invalid data, and yellow area depicts the same area without them. The green area shows the final state of the data.

physiological stress states in real-life situations, drawing from the knowledge established in the laboratory setting. Furthermore, unlike existing approaches, *SOSW* incorporates handling of rich contextual information derived from real-life settings in addition to physiological data captured by smartwatches. This approach aims to further enhance stress detection accuracy in the wild. In the following sections, we provide a detailed explanation of the data processing pipelines used in our laboratory and field studies.

A. Laboratory data processing pipeline

As depicted in Fig. 4, our pipeline for detecting physiological stress under laboratory conditions consists of four main processes: 1) data extraction, 2) data preprocessing, 3) feature extraction, and 4) ML model generation and validation. In what follows, we provide details on these processes.

Data extraction: The data encompasses physiological measurements collected through commodity smartwatches during the laboratory study, as well as information about participants such as their pseudo-identity and age. Additionally, it contains logs detailing the participation process in the laboratory study protocol, including the start and end timestamps for each laboratory scenario. The laboratory study protocol was designed to induce stress while recording physiological data from smartwatch PPG and chest strap ECG sensors[3]. More details about the data collection and laboratory study protocol can be found in Sec. V.

Given that *SOSW* primarily relies on smartwatch measurements, we extract raw PPG signals the smartwatch and their corresponding stress labels. The physiological readings were matched with corresponding ground truth labels using their timestamps and the start and end timestamps of laboratory study scenarios. These raw data are subsequently subjected to a comprehensive data preprocessing procedures to yield refined datasets.

Data preprocessing: Rigorous data preprocessing methodology is essential to ensure that the PPG data from a commercial smartwatch is reliable. Therefore, we paid special attention to our preprocessing methods, which comprise the following five sub-processes: band-pass filtering, outlier removal, interpolation, peak detection, and handling invalid HR and IBI data. The first four sub-processes are illustrated in Fig. 5, and the last (fifth) sub-process is illustrated in Fig. 6. We use these five sub-processes to get the most accurate HR and IBI measurements from commodity smartwatch PPG signals. It is worth mentioning that we employ the IBI over HRV due to its simpler acquisition and a higher granularity of information. Although the HRV data may also be employed, it is important to note that the HRV is a derivative metric that is entirely dependent on the IBI data. Therefore, while HRV provides a view of autonomic nervous system activity and its modulation of cardiac function, the IBI data can yield more immediate and specific insights into cardiac rhythms, particularly in response to acute stressors.

Band-pass filtering: First, we filter the BVP signals by employing a third-order Butterworth band-pass filter to extract the heartbeat information with lower and upper cutoff frequencies of 0.5 Hz and 3.7 Hz, respectively [9], [50]. By setting the cutoff frequencies to such values, we guarantee that undesired components are eliminated, only preserving the relevant frequencies associated with the heartbeat.

Outlier removal: After the BVP signal is filtered, the signal passes through the outlier removal process. The outliers in the BVP signal can arise from abrupt shifts in sensor positioning on the skin or can be attributed to technical anomalies related to the sensor or the smartwatch's operating system. The outlier removal process employs a robust statistical technique

[3] The reliable HR data of chest strap ECG is used as ground truth to validate the HR data estimated from the smartwatch PPG sensor readings.

known as the 4×MAD (Median Absolute Deviations) method [51]. The 4×MAD is an efficient method for identifying and eliminating sharp, sudden spikes within the BVP signal. By applying this method, the outliers in the BVP signal are carefully removed, ensuring that the resultant signal retains more reliable and genuine information related to heart activity.

Interpolation: Due to the removal of the outliers, the BVP signal becomes non-continuous. To address this issue, linear interpolation is employed. Although there exist other interpolation techniques [52] such as full degree polynomial interpolation, piecewise cubic Hermite interpolation, we selected the linear interpolation due to its robustness to noise, straightforwardness, and computational efficiency.

Peak detection: The interpolated signal passes through the peak detection process in which the signal is normalized in the range (0,1) first before finding the peaks. To find the peaks and ultimately estimate the HR and IBI, the heartpy open-source Python library [53] is employed.

Handling invalid HR and IBI data: Fig. 6 depicts the final subprocess of our data preprocessing pipeline, handling of invalid HR and IBI data. There might be unavoidable cases in which the estimation of the HR and IBI are invalid due to the noise of the BVP readings in the wild from the smartwatch. In these cases, to guarantee appropriate values of HR and IBI, invalid values are discarded. Consequently, estimated HR values that fall outside the physiologically plausible range of 30 to 220 beats per minute [9], [50] are replaced with 'NaN' to denote their invalidity. As a consequence, non-continuous data (with gaps) are created. To solve this problem, in instances where IBI and HR values are missing, we apply linear interpolation which offers a reasonable estimation for continuity.

Feature extraction: The estimated HR and IBI are now utilized to generate a pool of time-domain physiological stress features. These features, which have demonstrated efficacy in detecting physiological stress [9], [11], encompass various statistical metrics derived from the HR and IBI data. These features include the minimum, maximum, mean, median, standard deviation, kurtosis, skewness, 20-th and 80-th percentiles of the HR and IBI data, alongside the standard deviation of successive differences (SDSD) and root mean square of SD (RMSSD) between consecutive IBIs. To create these features a sliding window of 1-minute is applied to both the HR and IBI time series data. It is worth mentioning that the 1-minute window size is considered a standard for laboratory and ambulatory physiological monitoring [9], [29], [54], [55], [56].

ML model generation and validation: Several ML learning models such as the adaptive boosting (AdaBoost), gradient boosting (GB), logistic regression (LR), multilayer perception (MLP), random forest (RF), support vector machine (SVM), and extreme gradient boosting (XGB) are evaluated in order to select a model with the best performance. To train and evaluate the ML models, we have employed the leave-one-subject-out cross-validation (LOSO CV) which is a robust evaluation technique commonly used for testing the generalizability of an ML model in human studies [26], [57]. We evaluate our models with LOSO CV technique on the entire dataset (features, labels) obtained from 26 participants. We complete our laboratory data processing pipeline by developing a precise physiological stress model from the entire laboratory dataset. This model is subsequently applied in the "lab knowledge transfer" step of field data processing pipeline, which we elaborate in the following subsection.

B. Field data processing pipeline

The objective of the field data processing pipeline is to detect perceived stress under real-life conditions. To achieve this, our pipeline utilizes physiological data from smartwatches and contextual information from smartphones. We meticulously preprocess this data in a systematic and comprehensive manner, applying laboratory-based knowledge to real-world scenarios. The pipeline, illustrated in Fig. 7, addresses the motion and looseness artifacts associated with commodity smartwatches in realistic conditions. This ensures the reliability of HR and IBI data. Additionally, the pipeline integrates diverse situational context data from passive sensing and EMA. Similar to the laboratory data processing pipeline, our field data processing pipeline consists of four key steps: 1) data extraction, 2) data preprocessing, 3) feature extraction, and 4) ML model generation and validation. The following paragraphs elaborate on each of these steps.

Data extraction: We extract the physiological and wrist-motion data gathered through commodity smartwatches, along with passive sensing-based contextual data and EMA-based contextual data collected using participants' smartphones. The data also encompasses participants' pseudo-identities and perceived stress data collected via EMA, which is utilized to create binary stress labels in real-world scenarios.

From the smartwatch, we extract raw PPG and tri-axial acceleration signals. Simultaneously, from the smartphone, we extract users' perceived stress using EMA, as well as their contextual information through both EMA and passive-sensing data. The EMA-based contextual data, as illustrated in Table III, includes such information as users' perceived stress, ongoing activities, location, and social settings. Additionally, the timestamps of EMA are leveraged to derive further contextual information, such as the day of the week and the hour of the day. Regarding passive-sensing-based contextual data, as presented in Table II, we extract information related to activity recognition, transitions, call logs, device screen state, and location. Finally, the extracted physiological and wrist-motion data from the smartwatch, along with the contextual data from the smartphone, are matched with corresponding ground truth labels.

The self-reported PSS-4 data [58] from EMA were utilized to infer binary stress labels (*i.e.,* ground truth) during the

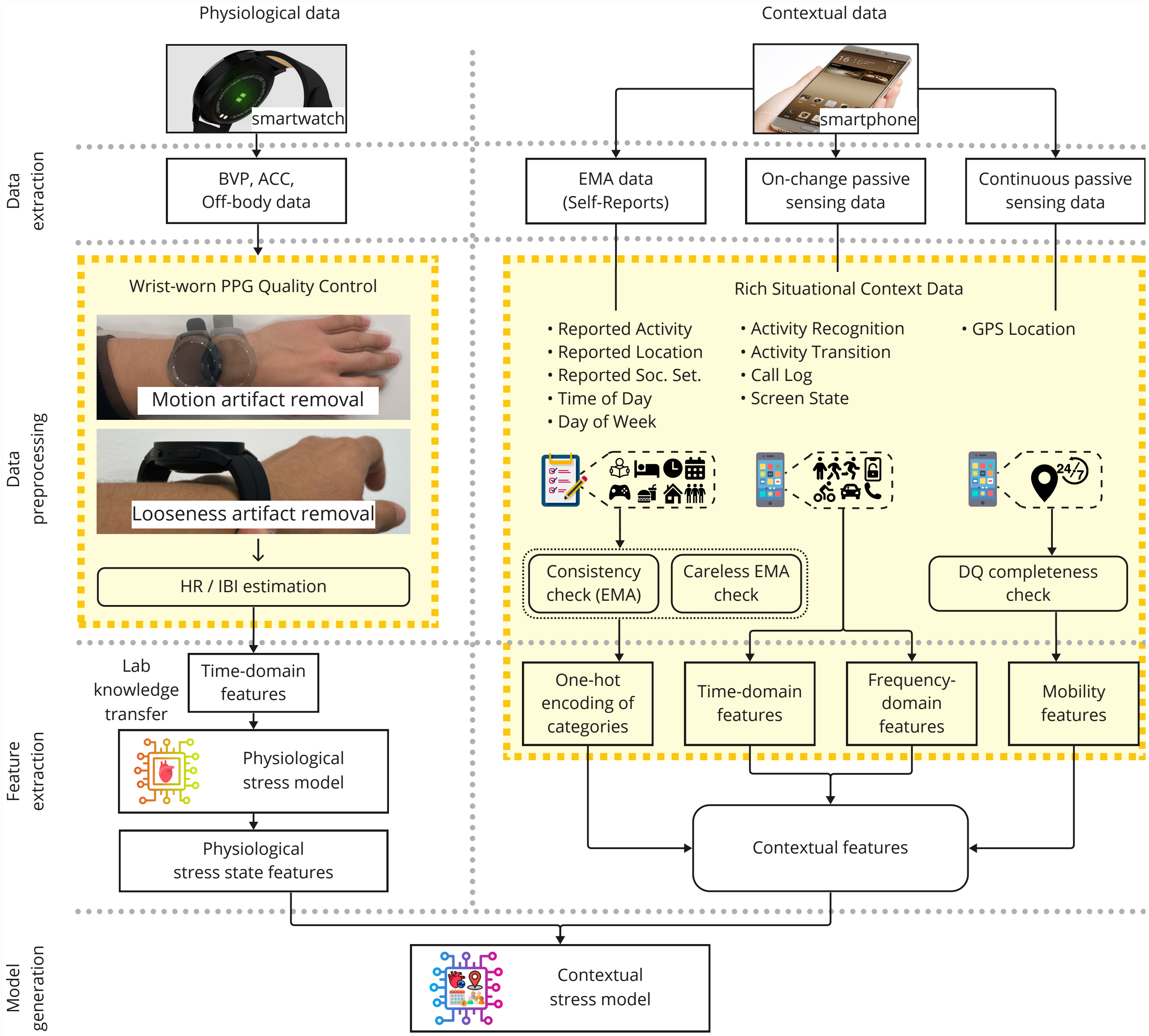

Fig. 7: Field data processing pipeline involving PPG signals of commodity smartwatch and contextual data of smartphone.

training and validation steps. The two positive items, specifically questions 2 and 3 in Table III, are reverse coded, and the average of all stress-related questions is calculated for each participant. This averaged stress score serves as the threshold for subjective stress classification, with scores above and below or equal to the mean labeled as 'stressed' and 'not stressed,' respectively.

Data preprocessing: The preprocessing steps for the smartwatch PPG data largely follow those used in the laboratory stress model pipeline. However, a comprehensive quality control process called "Wrist-worn PPG Quality Control" is included for discarding unreliable PPG data and consequently improving the overall data quality. This process consists of three sub-processes: a) motion artifact removal, b) looseness artifact removal, and c) HR and IBI estimation.

Motion artifact removal: It has been argued that physiological arousal, that should be indicative of a stress response, can be easily obfuscated by activity confounds such as changes in posture, movement of hands, and physical activities (e.g., walking and running) [11]. Screening out such physical activity confounds can be a plausible method to reduce confusion on stress-sensing. Therefore, the motion artifact removal process identifies and removes specific segments of the BVP signals aroused by the activity confounds.

Fig. 8 visually illustrates the sequential procedures employed to eliminate the corrupted BVP signal caused by activity confounds. The tri-axial accelerometer values are

This article has been accepted for publication in IEEE Internet of Things Journal. 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/JIOT.2024.3375299

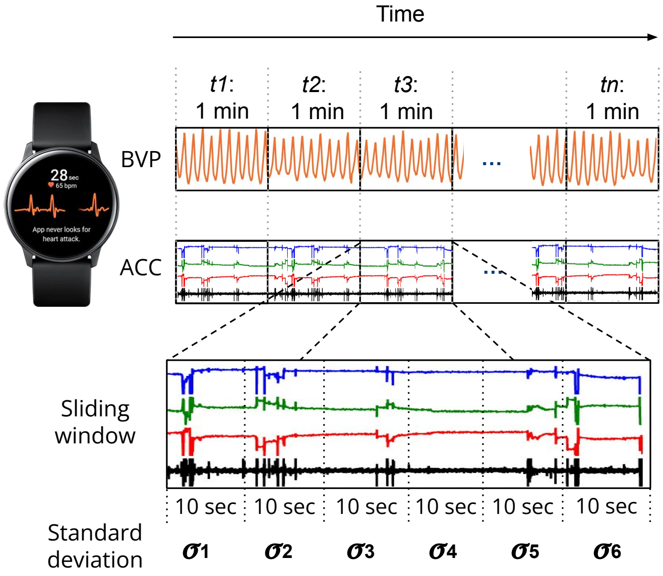

Fig. 8: Physical activity confound detection based on standard deviations of acceleration signals from commodity smartwatch.

utilized by analyzing the acceleration signal during a 1-minute time frame. To initiate the process, the timestamps of the BVP and acceleration signals are meticulously synchronized. Then, the one-minute interval of acceleration signal is partitioned into six smaller sub-intervals, each spanning a duration of 10 seconds. The standard deviation (denoted with σ) of the acceleration magnitude is computed from each of the six sub-windows. If the standard deviation of at least three sub-windows exceeds a threshold value, the associated 1-minute window in the BVP signal is considered to be compromised and thus excluded from further processing. Here, the threshold value of 0.21384 was chosen based on empirical evidence, which demonstrated that this particular threshold allows for the accurate detection of both static and non-stationary states [59].

Looseness artifact removal: We have shown in our preliminary study that BVP signal readings are considerably affected by the degrees of loose wearing conditions of smartwatch, see Sec. III. Therefore, to filter out the loose wearing conditions and consequently obtain cleaner BVP signals, the variance of light intensity readings from the smartwatch can be utilized [18]. Particularly, we employ the hidden Markov model (HMM) based on the Viterbi algorithm [18]. The HMM comprises two hidden states (accurate and inaccurate) and learns the transitions and emission probabilities based on observations. The HMM is trained on the variance of the PPG light intensity. To optimize the training process and enhance its efficiency, the Baum-Welch expectation-maximization algorithm is employed. Employing these methods not only ensures that the measurements are free from the interference of looseness artifacts but also guarantees a higher degree of accuracy and precision in the results derived from the PPG light intensity data [21].

TABLE II: Summary of passive sensing-based contextual data collected in the wild using Android smartphones.

Data	Values
Activity recognition	`activity_type`, `confidence`
Activity transition	`activity_type`, `transition`
Call log	`timestamp`, `call_type`, `duration`
Device screen state	`screen_state`, `key_restriction`
GPS location	`latitude`, `longitude`, `accuracy`

Following the processes for motion artifact removal and looseness artifact removal, additional preprocessing steps are applied to the BVP signal. These steps include band-pass filtering, outlier removal, interpolation, peak detection, and handling of invalid data. These preprocessing steps are same as those described in Sec. IV-A. These additional preprocessing steps assist in obtaining more accurate HR and IBI values, thereby help in deriving more reliable physiological stress features.

Feature extraction: Since the *SOSW* field pipeline follows a two-layer learning architecture, we first extract features for the ML model in the first layer. Similar to the case of the feature extraction of the laboratory stress data processing pipeline described in Sec. IV-A, we derive 1 minute-level time-domain features from the estimated HR and IBI data, such as minimum, maximum, standard deviation, kurtosis, among other features.

Then, we feed these features to our ML model in the first layer, the laboratory-based physiological model created in Sec. IV-A, to obtain a refined feature, *i.e.,* the likelihood of a person being stressed. It is worth mentioning that to create this refined feature obtained from the smartwatch, a knowledge transfer type was devised, which is a well-trained and validated ML model with ideal stress and non-stress dataset was exploited. Provided these 20 time-domain features, this ML model learned to distinguish between stress and non-stress of a person providing a probabilistic distinction, which is used as a part of our pool of features for our ML model in the second layer.

As a part of the pool of features for our ML model in the second layer, we incorporate features extracted from the smartphone passive sensing data related to participants' physical activities, device use patterns, and mobility behavior. These passive sensing-based features include time and frequency domain information derived from the activity recognition, transition, call log, screen state, and GPS location data. Furthermore, we generate additional features from the user-reported contextual data (*i.e.,* EMA), including information about participants' ongoing activities, their locations, and social settings. To create these features, we perform one-hot encoding on the responses, converting the categorical data of raw EMA responses into sets of binary encoded values. These features are then employed to generate our field stress model.

TABLE III: Summary of EMA-based contextual and perceived stress data collected in the wild using smartphones.

EMA Prompt	Answer options
Perceived stress	
1. How often have you felt that you were unable to control the important things? 2. How often have you felt confident about your ability to handle problems? 3. How often did you feel that things were going your way? 4. How often did you feel that difficulties were piling up so high that you could not overcome them?	5-point Likert scale (Never - Very often)
Context	
5. Please report your current activity.	Working or studying, sleeping, resting or relaxing, video watching, class or meeting, eating or drinking, gaming, conversing, getting ready for bed, calling or texting, right after waking up, driving, other activity
6. Please report your current location.	Home, work, restaurant, vehicle, other
7. Please report your current social settings.	Social, asocial

ML model generation and validation: Similar to the model generation and validation in the laboratory stress process pipeline in Sec. IV-A, our second (final) layer ML model is generated and rigorously validated using the LOSO CV. Each participant is iteratively left out as an 'unseen' participant for testing purposes, while the model is trained on the rest of the participants. This process is repeated for each participant in the dataset, allowing for a comprehensive assessment of the model's performance across different users.

V. Data collection

In this section, we provide a comprehensive overview of our data collection methodology, detailing the types of data collected, devices and applications used for data acquisition, as well as the participant recruitment and demographic information. We also outline the study procedures that governed data collection, ensuring transparency in our approach of acquiring the necessary information for our stress detection model.

A. Data types

Physiological data: We recorded raw PPG signals to monitor user's heart activity, and tri-axial accelerometer readings to track user's wrist motion, for which we utilized off-the-shelf smartwatch, *i.e.,* Galaxy Watch 5. The data from the smartwatch were captured at a sampling rate of 12 Hz for both PPG and accelerometer signals during laboratory and field studies. Additionally, during the laboratory study exclusively, we collected HR and IBI data at 1Hz sampling rate using a Polar H10 chest strap [48], which is noted for its accuracy in previous studies [49].

Contextual data: Alongside physiological and wrist motion data of smartwatch, we also collected rich contextual data through EMA and passive sensing techniques using smartphone. The rich contextual data provides insights that complement the physiological data captured by the smartwatch, allowing for a greater representation of the experiences of the participants in real-life scenarios.

The EMA data comprises participants' perceived stress state and their situational contexts. Table III shows the EMA questions with their possible answers prompted to the participants. User's perceived stress state was self-evaluated by participants with PSS-4 questionnaire [58], which is widely used in prior works [11], [60]. Concurrently, we adopt the method proposed in [61] to capture user's situational context information through EMA. The method includes 13 categories for user's activity, 5 categories for location, and a binary social settings, all of which represent the user's in-situ context at the time of filling out the EMA. Moreover, we devise an additional contextual data from the time of filling out the EMA, namely the time of day and day of week.

In addition to the EMA data, we expanded our set of contextual data by incorporating passive sensing data of smartphone. The summary of the incorporated passive sensing data types are provided in Table II. The data included user's activity information recognized by smartphones provided by activity recognition API of the Android operating system (OS). This API provides updates when a user transitions between activities, for example, from walking to being still. Furthermore, it periodically reports detected activities, accompanied by a confidence rating that indicates, for example, if the device is with a user who is walking and the probability that this activity has been correctly identified.

Alongside activity data, we also collected device use information, such as phone calls and device screen state. Phone call data includes such information as the time each call occurred, whether it was an incoming, outgoing, or missed call, and how long the call lasted in seconds. On the other hand, the device screen state information includes the state of the smartphone's screen ('on' or 'off') and keyguard (virtual 'lock') state, which collected whenever user interacts with the smartphone screen, *i.e.,* on-change. For example, data is recorded each time the user unlocks the screen or turns it off.

We also collected GPS location data, including latitude and longitude expressed in degrees, accompanied by information about the location's accuracy measured in meters. The accu-

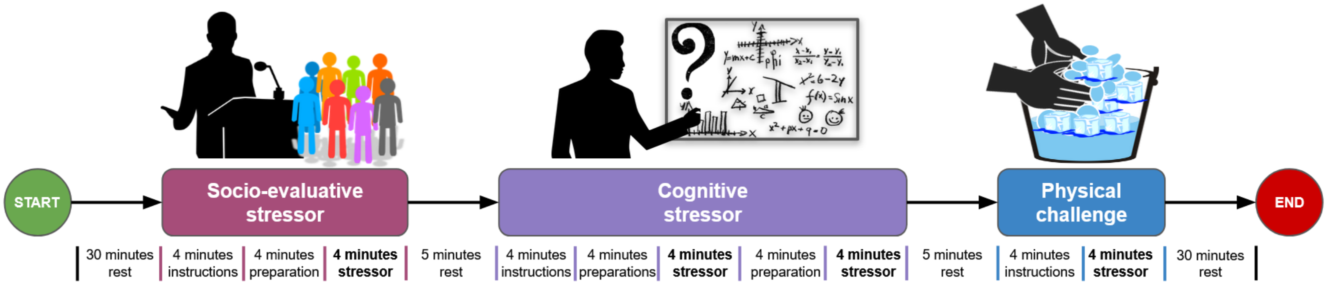

Fig. 9: Three-stage laboratory stress study protocol, including socio-evaluative stress by public speech, cognitive stress by time-constrained mental arithmetic, and physical challenge by cold-pressor test.

racy value signifies the radius of the possible area where the user could be located, with smaller values indicating more precise and accurate location information.

B. Data collection devices

We utilized a strategic combination of devices to capture physiological and contextual data during both laboratory and field stress studies. For both lab and field evaluation, we used the Samsung Watch 5 smartwatch [47], aiming to determine the viability of using commodity smartwatches for stress detection. The choice of this particular device was driven by two main factors: 1) unlike most of the commercially available smartwatches such as Fitbit by Google and Apple Watch, the Galaxy Watch allows access to the raw sensing data [62], and 2) its widespread popularity that makes it suitable for practical applications [36], [63].

During the laboratory phase, HR data using the Polar H10 chest strap were additionally gathered. In the field study, while still leveraging the Samsung Watch 5 for physiological readings and wrist-motion data, we employed Android smartphones to gather EMA and passive sensing data. The utilization of these devices allowed for a comprehensive exploration of stress behaviors in real-life scenarios.

C. Data collection applications

EMA and passive sensing data collection app: We developed a custom Android application to acquire EMA and passive sensing data from users' smartphones. With regards to EMA data collection, we integrated firebase cloud messaging [64] to trigger push notifications on user smartphones whenever they are required to submit EMA self-reports. And for passive sensing data collection, we deployed an always-running foreground service in the smartphone, guaranteeing uninterrupted acquisition of smartphone passive sensing data. Lastly, we carefully refined the application's user interface (UI) to enhance the user experience, with the goal of achieving high user compliance.

Physiological and wrist-motion data collection app: We also developed a custom application for WearOS smartwatch OS, and similar to the smartphone application, it was also deployed as an always-running foreground service. The application also provided a watchface UI, in order to provide essential information at a glance (e.g., current date and time), and to consistently keep the application in the foreground mode. Additionally, we adjusted specific system configurations of smartwatches to facilitate uninterrupted, continuous data collection by our application. These adjustments included disabling the battery-saving 'doze' mode and minimizing interruptions from the watch's native functions, such as its activity and sleep detection features.

Data collection server: On the data collection server side, we integrated our mobile data collector with the Easytrack data collection platform [65], ensuring efficient data transmissions between participants' devices and the server through the gRPC framework [66]. The Easytrack platform features general-purpose functionalities that enabled easy adaptation to our stress detection use-case. Easytrack's server regularly computes various data quality metrics and reports them to the researchers, simplifying the monitoring process and ensuring data integrity. A dashboard by EashTrack provides an overview of data quality, enabling quick identification of any potential issues or discrepancies about each participant and each source of data (e.g., sensor or EMA).

D. Participant recruitment and demography

For our study, we recruited participants using email broadcasts, flyers at a local university, and announcements on local social media platforms, specifically targeting Android device users. We successfully recruited 28 healthy participants, comprising 16 university students and 12 from the general population. Of these participants, 15 were female and 13 were male. The average age of the participants was 22.8 years, with a standard deviation of 2.7 years.

E. Data collection procedure

Laboratory data collection procedure: Participants in the laboratory data collection study underwent three distinct, validated stress-inducing scenarios, in line with earlier studies [9], [29]. These stressors encompassed socio-evaluative, cognitive, and physical challenges. The sequence and duration of these stressors, as well as the initial 30-minute baseline resting period, are illustrated in Fig. 9. During the baseline rest period, participants were instructed to sit comfortably in an empty room and relax as much as possible, with controlled room temperature and lighting to prevent external factors cause any unintended physiological arousal. Additionally, participants were asked to refrain from using their smartphones throughout the laboratory study.

After the initial 30-minute baseline rest period, participants engaged in three stress-inducing scenarios.

- *Socio-evaluative stressor*: Participants were given 4 minutes to prepare before delivering public speech in front of five researchers, including a professor. To intensify the social stress associated with the task, their speeches were video recorded, ensuring anonymity. After the speech, participants had 5 minutes of rest to recover.
- *Cognitive stressor*: Participants underwent two sessions of 4-minute cognitive stressors: one while sitting on a chair and the other in a standing position. The stressor included a mental arithmetic task, specifically counting backwards in steps of 7, similar to the method used in [9]. To increase arousal, participants were shown their progress in real-time, and rewards were promised for top performers. Five minutes of rest followed this task.
- *Physical challenge*: Last stressor was the cold pressor test, where participants were instructed to immerse their hands in ice-cold water for up to 4 minutes. Most participants lasted around two minutes. Following this, participants had a 30-minute resting period to conclude the experiment.

Field data collection procedure: The two-week field data collection was performed in unconstrained natural environments where participants were not subjected to any predetermined protocol. Only minimal guidelines to ensure the uninterrupted collection of data, and to maintain the study's integrity, were instructed to the participants. They were instructed to wear the smartwatch continuously, except during bedtime or if it caused skin irritations. Participants were also reminded to keep data collection applications on their smartphones and smartwatches operational, with a particular emphasis on keeping the smartwatch application in the foreground to avoid data collection stoppage. Participants were encouraged to keep sensor permissions active, including GPS, to collect location data.

To effectively capture the perceived stress dynamics throughout the day, participants received 12 push notifications daily, prompting them to complete the EMA questionnaire at randomized intervals between 40 to 80 minutes, allowing for a comprehensive understanding of stress experiences in various situations. The questionnaire was designed with precision to minimize ambiguities, with a prompt asking participants to reflect exclusively on the preceding hour to enhance accuracy of the response and reduce memory biases, effectively capturing the real-time stress experiences.

VI. Evaluation

This section presents a thorough evaluation of our proposed methodology, *SOSW*, with its performance in detecting stress under laboratory and real-life conditions. We start with an overview of our dataset and then move to a detailed explanation of our evaluation methods. Following this, we present the results from evaluations conducted on both laboratory and field study datasets.

A. Dataset summary

We present a detailed dataset summary, providing insights into our data collection and cleansing processes. For clarity, we have chosen to report the amounts of data in minutes for our laboratory study dataset and in hours for our field study dataset.

In the laboratory study dataset, we initially recruited 28 participants, but two participants withdrew from the laboratory study, leaving us with a total of 26 participants. Each participant contributed 2 hours of data during the data collection phase. The dataset comprises 864 minutes of baseline-rest data and 323 minutes of stressor-related data, offering valuable insights into physiological responses to various stressors.

In our field study dataset, we collected data from 28 participants over a period of 2 weeks. This included a total of 1928 hours of physiological and wrist-motion data, 4930 hours of passive sensing contextual data, and 2867 hours of EMA contextual data with a total of 3732 EMA responses. To ensure the high quality of the data, we meticulously cleansed the dataset by removing participants and segments with invalid or missing data. Specifically, two individuals who had limited participation in our laboratory study were excluded from the field study dataset. Their inclusion could have introduced unwanted noise or biases, hence their exclusion.

In our field study, EMA data provided essential information about participants' perceived stress in uncontrolled, real-life settings. The EMA data was a crucial part of the field dataset, hence, participant compliance was highly important. However, two participants demonstrated significantly lower EMA compliance rate, submitting significantly less EMA data over two weeks compared to the rest of the group, leading to their exclusion from the field dataset. Furthermore, six participants exhibited a pattern of inputting similar EMA responses throughout the entire 2-week data collection period. A lack of variations in self-report data raised concerns about reliability and bias, and we decided to exclude these six participants from the field study dataset.

Further data cleansing steps involved excluding segments with missing or invalid data, considering the continuous sensors and their expected sampling rates. After this step, we were left with 651 hours of complete and usable data. Additionally, we applied our motion and looseness artifact removal filters, which resulted in the exclusion of 26 and 24 hours of data, respectively. The final dataset, therefore, contains 601 hours of high-quality data (with 727 remaining

TABLE IV: Statistical significance analysis of physiological stress features in laboratory study dataset using Welch's t-test. Table reports the significance of physiological features derived from commodity smartwatch data in distinguishing between baseline rest and three stressors.

Feature	Socio-Eval. Stressor		Cognitive Stressor		Physical Challenge	
	t-Stat	p-Value	t-Stat	p-Value	t-Stat	p-Value
Min HR	**-3.690**	**<0.001**	**-10.365**	**<0.001**	**-6.425**	**<0.001**
Max HR	**-10.752**	**<0.001**	**-10.332**	**<0.001**	-0.449	0.654
Mean HR	**-14.406**	**<0.001**	**-14.728**	**<0.001**	**-3.912**	**<0.001**
Median HR	**-14.301**	**<0.001**	**-14.632**	**<0.001**	**-4.493**	**<0.001**
Std of HR	**-7.520**	**<0.001**	**-5.504**	**<0.001**	**2.276**	**0.024**
Kurtosis of HR	-1.561	0.120	-1.844	0.068	0.214	0.830
Skewness of HR	-0.628	0.531	-0.859	0.392	0.543	0.588
20th Percentile HR	**-14.618**	**<0.001**	**-14.968**	**<0.001**	**-5.998**	**<0.001**
80th Percentile HR	**-11.923**	**<0.001**	**-11.716**	**<0.001**	-1.604	0.110
Min IBI	**13.815**	**<0.001**	**13.104**	**<0.001**	1.974	0.050
Max IBI	**3.093**	**0.002**	**10.065**	**<0.001**	**6.860**	**<0.001**
Mean IBI	**17.957**	**<0.001**	**17.344**	**<0.001**	**5.600**	**<0.001**
Median IBI	**17.514**	**<0.001**	**17.020**	**<0.001**	**5.700**	**<0.001**
Std of IBI	**-6.444**	**<0.001**	**-2.775**	**0.006**	**2.562**	**0.011**
SDSD	**-8.425**	**<0.001**	**-5.169**	**<0.001**	1.947	0.053
Kurtosis of IBI	**-2.023**	**0.045**	-1.052	0.294	0.583	0.561
Skewness of IBI	**-2.904**	**0.004**	-1.359	0.176	-0.386	0.700
20th Percentile IBI	**15.216**	**<0.001**	**14.450**	**<0.001**	**3.442**	**<0.001**
80th Percentile IBI	**17.552**	**<0.001**	**16.870**	**<0.001**	**6.613**	**<0.001**
RMSSD	**-8.469**	**<0.001**	**-5.129**	**<0.001**	1.965	0.051

EMA responses), after the exclusion of participants and segments with invalid or unreliable data, ensuring that our dataset is of the highest quality for analysis and research.

B. Evaluation methods

In this section, we describe the evaluation methods and performance metrics used for assessing our laboratory and field study datasets. Through this, our objective is to examine the generalizability of our work.

We start by assessing if features extracted from commodity smartwatch PPG signals can distinguish between resting and stress-induced periods in the laboratory study. For this, we use Welch's t-test of unequal variances to identify statistically significant differences between two binary stress groups: stress and non-stress. Similarly, we apply the same statistical test to evaluate the significance of features from passive sensing-based contextual data in the field study. For EMA-based contextual features, we utilize one-way ANOVA tests. These tests help us investigate if categorical group differences can statistically explain the variance in the overall dataset regarding perceived stress levels. By employing these methods, we gain preliminary insights into the reliability of these features, which aids in the development of machine learning models.

To evaluate the generalizability of our laboratory and field models, we apply LOSO CV technique. It is an effective technique for testing the robustness of our methods by systematically leaving out one subject's data at a time, helping us ensure that our models can be applied to the data of unseen users. In LOSO CV, each participant's data is treated as a single isolated fold, ensuring the model is tested against each individual-specific variations in the data. We employ widely used performance metrics, such as precision, recall, F-1 score, specificity, accuracy, and AUROC (Area Under the Receiver Operating Characteristic Curve). In summary, we aggregate these metrics from LOSO CV, combining results obtained after testing against each participant to provide a comprehensive assessment.

C. Predicting stress based on the laboratory dataset

This section focuses on stress detection model performance analysis in ideal, controlled laboratory settings. Evaluation in these settings can provide insights into the reliability of physiological data from commodity smartwatch in detecting stress under ideal circumstances. First, we investigate the distinctive physiological stress features, and then we leverage such features to explore the performance of various ML models in classifying stress in laboratory settings. To evaluate the effectiveness of our methodology, we also conduct a benchmark comparison against recent stress detection methods. In addition to smartwatch-based stress model performance, we also report our ECG-based laboratory stress classification performance as well.

Significant physiological features: We begin by determining whether the extracted features from the BVP signals are able to capture a significance difference between the resting and stress-induced periods of the laboratory study. To this end, we employ the Welch's t-test of unequal variances to determine which features showed any statistically significant differences between the resting baseline period and each of the stress-induction periods.

Table IV presents the results of the Welch's t-test, examining the significance of various physiological features extracted from PPG sensor data for distinguishing between resting and stress-induced periods for three different stressors: socio-evaluative stressor, cognitive stressor, and physi-

TABLE V: Comparison of performances of several ML methods on the task of binary stress classification using physiological data from laboratory study using commodity smartwatch data.

ML Methods	Precision	Recall	F-1 score	Specificity	Accuracy	AUROC
AdaBoost	0.812	0.777	0.794	0.746	0.764	0.761
GB	0.824	0.773	0.798	0.776	0.775	0.775
LR	**0.848**	**0.822**	**0.835**	**0.808**	**0.816**	**0.815**
MLP	0.826	0.815	0.820	0.772	0.796	0.794
RF	0.830	0.776	0.802	0.786	0.780	0.781
SVM	0.831	0.779	0.804	0.781	0.780	0.780
XGBoost	0.839	0.789	0.813	0.794	0.792	0.792

cal challenge. Several features exhibit statistical significance across all stressors, indicating their effectiveness in capturing physiological changes associated with stress. However, the physical challenge shows fewer significant features compared to the other two stressors. This suggests that the physical challenge may be less effective in inducing stress across all participants. In contrast, the socio-evaluative stressor emerges as the most effective stressor for inducing stress response. These results emphasize the potential of using these features to detect stress-related physiological changes using PPG measurements. This suggests that PPG-based stress detection approaches may yield comparable results to stress detection approaches employing ECG measurements.

Comparison of ML methods: Having determined that the features extracted from HR data (obtained from commercially available, off-the-shelf smartwatches) showed significant differences between rest and stress-induced periods, we then used these features to build ML models designed to infer whether the person is stressed or not stressed. The ML models considered here are the adaptive boosting (AdaBoost), gradient boosting (GB), logistic regression (LR), multilayer perception (MLP), random forest (RF), support vector machine (SVM), and extreme gradient boosting (XGBoost).

Table V shows the performance comparison of these ML models, where several metrics including F-1 score, precision, recall, specificity, and so on, are presented. Among these metrics, F-1 score is a critical metric that balances precision and recall, offering a comprehensive view of the performance of the respective models. The LR stands out as the top-performing model with an F-1 score of 0.835, indicating its strong ability to accurately identify individuals experiencing stress while maintaining a low rate of false positives. Although other models like XGBoost and MLP also perform well, the LR model demonstrates its proficiency in the binary stress classification, making it a compelling choice. Additionally, it's worth noting that the LR model consistently exhibits superior performance across all the other performance metrics compared to other models.

Benchmark comparison: To assess the stress detection performance of *SOSW* in laboratory settings, we conduct a comparative analysis that includes a recent approach developed by Dai et al. [28] utilizing PPG-equipped wristband, and the work by Mishra et al. [9] that utilizes ECG sensor. In this performance evaluation, we consider the best-performing models from each of these studies. The comparison is based on three commonly used performance metrics: precision, recall, and F-1 score.

Fig. 10 shows the performance of several works for detecting stress under laboratory conditions. This figure shows that the proposed *SOSW* provides the best performance among the works using physiological measurements obtained from the commodity smartwatch PPG sensor. The *SOSW* provides an F-1 score metric value of approximately 0.84 while the approach proposed by Dai et al. 0.62 [28]. Additionally, when compared to the approach by Mishra et al. [9], *SOSW* shows notable proximity in the precision metric with a value of 0.85 as opposed to the prior work's reported 0.86 precision. Interestingly, it is also observed that when *SOSW* employs the ECG measurements, it provides higher F-1 score than when it uses PPG measurements. This improved performance can likely be attributed to the generally higher accuracy of electrical-based ECG sensors, as opposed to the optical-based PPG sensors. Moreover, although the *SOSW* yields a marginally diminished recall, it exhibits superior precision and near similar recall in comparison to the approach proposed by Mishra et al. [9]. These results demonstrate the robustness of our methodology in detecting stress in laboratory settings and affirm the potential of commodity smartwatches in physiological stress detection.

D. Predicting stress based on the field dataset

This section evaluates the performance of the *SOSW* stress detection methodology in the wild. We begin by examining the impact of contextual factors on perceived stress levels. Then, we comprehensively evaluate diverse ML models on the challenging task of detecting stress in the wild. To determine the effectiveness of our approach, we also perform a benchmark comparison with previous research that uses dedicated sensors to obtain precise physiological readings in real-world settings.

Significant contextual features: We start our evaluations by assessing the significance of contextual features derived from EMA on the perceived stress levels. To this end, we conduct five separate one-way ANOVA tests with the perceived stress score as the dependent variable and the EMA reported activity, location, social setting, time of day, and day of week as independent categorical values, respectively. The results of these ANOVA tests are summarized in Table VII. This table

TABLE VI: Statistical significance analysis of passive-sensing-based contextual features in field study dataset using Welch's t-test. Table reports the significance of various contextual features in identifying users' (binary) perceived stress state.

Category	Subcategory	Feature	t-Stat	p-Value
Activity	Physical Activity	`unique_activities` duration	**3.560**	**<0.001**
		`on_foot` duration	**2.937**	**0.003**
		`on_foot` frequency	**2.641**	**0.008**
		`unknown` frequency	**2.494**	**0.013**
		`in_vehicle` number	**2.086**	**0.037**
		`unknown` duration	1.644	0.100
		`still` frequency	1.442	0.150
		`in_vehicle` frequency	1.233	0.218
		`on_bicycle` frequency	1.210	0.226
		`on_bicycle` duration	1.057	0.291
		`still` duration	0.278	0.781
	Mobility	`mean_distance`	**2.902**	**0.004**
		`std_distance`	1.880	0.060
		`mean_speed`	1.874	0.061
		`total_distance`	1.577	0.115
		`max_distance`	1.362	0.173
		`std_speed`	1.065	0.287
		`min_speed`	0.802	0.422
		`min_distance`	0.507	0.613
		`max_speed`	0.309	0.757
Device usage	Call log	`avg_outgoing_calls` duration	**-2.093**	**0.037**
		`max_outgoing_calls` duration	**-2.029**	**0.043**
		`min_outgoing_calls` duration	-1.948	0.052
		`total_outgoing_calls` duration	-1.824	0.068
		`std_incoming_calls` duration	1.493	0.136
		`total_incoming_calls` duration	0.799	0.424
		`unique_incoming_calls` number	0.779	0.436
		`max_incoming_calls` duration	0.721	0.471
		`avg_incoming_calls` duration	0.657	0.511
		`incoming_calls` frequency	0.652	0.514
		`min_incoming_calls` duration	0.505	0.614
		`outgoing_calls` number	-0.406	0.685
		`std_outgoing_calls` duration	-0.404	0.687
		`outgoing_calls` frequency	-0.289	0.773
		`unique_missed_calls` number	-0.206	0.837
		`missed_calls` frequency	0.202	0.840
	Screen state	`screen_on` duration	1.855	0.064
		`screen_off` frequency	1.426	0.154
		`screen_on` frequency	1.325	0.185
		`user_present` frequency	1.127	0.260
		`screen_off` duration	0.487	0.626
		`user_present` duration	0.345	0.730

TABLE VII: Statistical significance analysis of EMA-based contextual features from field study dataset using One-Way ANOVA test.

Feature category	F-statistic	p-Value
Reported activity	**5.573**	**<0.001**
Reported location	**13.816**	**<0.001**
Reported social settings	3.592	0.058
Hour of day	0.755	0.519
Day of week	**18.293**	**<0.001**

reveals that the categories related to reported activity, location, and day of week are statistically significant ($p < 0.001$). This indicates that there are significant changes in perceived stress levels linked with various activities, places, and days of the week (*i.e.,* weekday or weekend).

We also analyze the effect of time-domain and frequency-domain features extracted from passive sensing on the perceived stress levels, as shown in Table VI. Employing Welch's t-test, similar to our analysis of physiological features in Sec. VI-C, we find that several time-domain and frequency-domain features related to physical activity (passive sensing) data are statistically significant. However, device usage information, such as call logs, shows limited statistical significance. These findings suggest that our time-domain and frequency-domain contextual features, particularly those related to activity-based passive sensing data, have a correlation with perceived stress levels.

Comparison of ML methods: With these promising results, we further evaluate the performance of *SOSW* in the task detecting stress in the wild. The *SOSW* field data processing pipeline leverages a two-layer detection architecture, in which the first layer employs an accurate physiological stress model devised from the laboratory pipeline, specifically the LR model. This best-performing physiological stress model devised from the laboratory dataset is employed in the first

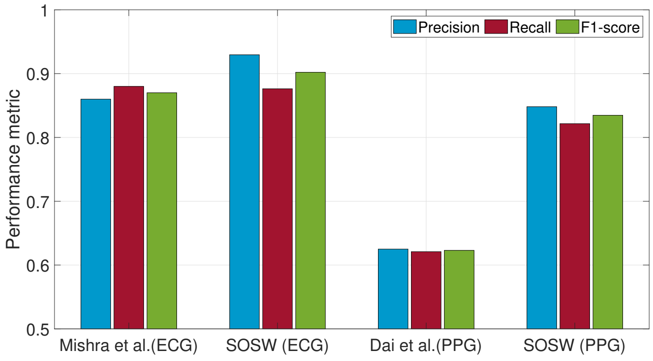

Fig. 10: Benchmarking of SOSW methodology against SOTA in detecting physiological stress in laboratory settings.

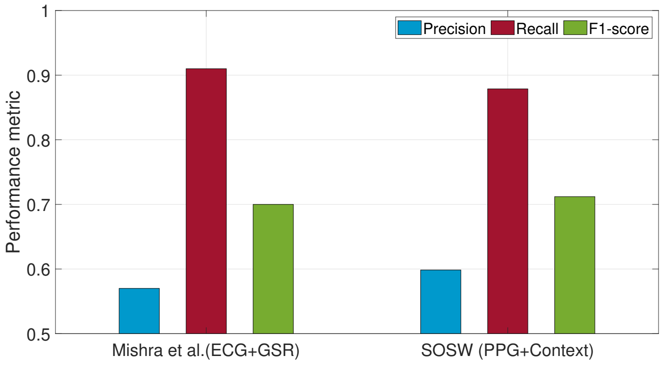

Fig. 11: Benchmarking of SOSW methodology against SOTA in detecting perceived stress in the wild.

layer of the field data processing pipeline. The second layer model is the contextual stress model, the final decision-maker in detecting perceived stress in the wild. We evaluate the efficacy of various machine learning (ML) models for this layer, including adaptive boosting, gradient boosting, logistic regression, multilayer perception, random forest, support vector machine, and extreme gradient boosting. Additionally, the effectiveness of these models is assessed across diverse combinations of data, including physiological, EMA, and passive sensing data.

Table VIII presents the performance analysis of the *SOSW* pipeline for stress detection with the field dataset using various ML models and different combinations of features, *i.e.,* variants. Several important insights can be drawn from this table. Firstly, the addition of contextual information to physiological data marginally enhances the model's accuracy. Second, fusing all features devised from field dataset results in an F-1 score of 0.681 with an increase of 2.3 percentage points compared to using only physiological data. Thirdly, the highest F-1 score in real-world conditions, at 0.712, is recorded when the previous stress state is considered. This aligns with finding from Mishra et al.[9], where authors report that accounting for previous stress state significantly enhances the final classification accuracy. Be that as it may, our findings suggest that a more extensive dataset might not always translate to better performance; in fact, it can potentially mislead the models in the classification task. Finally, among the various ML models, linear regression, gradient boosting, and support vector machine showed higher accuracies, while the multilayer perceptron achieved the maximum performance.

Benchmark comparison: To assess the stress detection performance of *SOSW* using the field study dataset, we conduct a comparative analysis with a previous work's findings reported in [9]. It is crucial to emphasize that the previous study operates with the advantage of high-precision ECG and GSR measurements acquired through dedicated hardware. In this performance evaluation, we consider the best-performing models from each of these studies. To be specific, for the *SOSW* the MLP model, which employs physiological data of general-purpose, commodity smartwatch and previous stress state is considered. The comparison is based on three commonly used performance metrics: precision, recall, and F-1 score.

Fig. 11 shows the performances of *SOSW* and the recent prior work by Mishra et al. [9] in the task of stress detection in the wild. Mishra et al. [9] reported achieving up to 0.70 F-1 score in their field study using ECG and GSR data. This figure shows that the contextual stress model by *SOSW* outperforms the counterpart by 1.2 percentage points in terms of F-1 score, utilizing commodity smartwatch and smartphone data. While *SOSW* records marginally lower recall of 0.88 compared to 0.91 by Mishra et al. [9], it gains the upper hand by outperforming in terms of precision with a value of 0.60, exceeding the 0.57 precision score of the prior work. And our methodology, *SOSW*, could achieve an F-1 score of 0.71 using commodity smartwatch physiological data and contextual information of smartphone. These findings highlight the potential of PPG data from commodity smartwatches, complemented by contextual information, as a viable alternative to traditional dedicated ECG and GSR sensors for stress detection in the wild.

VII. Discussion

A. Detecting stress in the lab and in the wild

Under laboratory conditions, it was found that physiological features extracted from off-the-shelf smartwatches are highly correlated with stress instances. By employing such features, several ML models were tested, and among them, the LR showed the best performance with an F-1 score of 0.835. We believe that such considerable improvement is achieved due to the fact that *SOSW* consists of a robust data processing pipeline and employs relevant features that are highly correlated with stress instances.

As stress detection research transitions from laboratory settings to real-life scenarios, researchers consistently report that

TABLE VIII: Comparison of several ML methods based on the F-1 score metric on the task of binary field stress classification using physiological and contextual features.

Variants	AdaBoost	GB	LR	MLP	RF	SVM	XGBoost
Physiological only	0.648	0.655	**0.658**	0.656	0.635	0.657	0.649
Physiological + Previous stress state	0.701	0.707	0.710	**0.712**	0.660	0.709	0.679
Physiological + EMA activity	0.647	**0.656**	0.652	0.648	0.642	0.647	0.642
Physiological + EMA location	0.647	0.655	**0.658**	0.654	0.632	0.648	0.643
Physiological + EMA social settings	0.647	**0.656**	0.652	0.654	0.636	0.655	0.650
Physiological + EMA hour of day	0.647	**0.659**	0.658	0.653	0.630	0.653	0.641
Physiological + EMA day of week	0.647	0.656	**0.660**	0.655	0.646	0.652	0.646
Physiological + Passive sensing activity	0.649	**0.654**	0.649	0.639	0.652	0.650	0.633
Physiological + Passive sensing call log	0.650	**0.651**	0.650	0.647	0.646	**0.651**	0.634
Physiological + Passive sensing screen state	0.647	0.653	**0.655**	0.649	0.649	0.647	0.634
Physiological + Passive sensing mobility	0.647	0.653	0.653	0.648	0.654	**0.661**	0.643
Physiological + All contextual features	0.666	0.675	**0.681**	0.618	0.665	0.674	0.652

contextual data improves the performance of stress detection [14], [22], contrary to solely relying on physiological signals. While dependence on an external data to the smartwatch could put limitations on the accuracy of the methodology, luckily, the modern smartwatches also have access to contextual data directly from the smartwatch itself. Such contextual data includes user's activities, step counts, sleep duration, and so on [67]. However, we leave the nuanced smartwatch-based contextual data, which has evolved fairly recently, for future studies. Instead, our work uses contextual data from smartphones, and smartphones have proven to be highly suitable for a variety of real-life scenarios and heterogeneous audiences [68].

In real-life conditions, *SOSW* not only exploits physiological features but features extracted from contextual information obtained through EMA and passive sensing. Additionally, *SOSW* exploits a type of knowledge transfer in which the well-trained and validated LR model is employed as a fixed model on its first layer. The second ML model was trained and validated employing contextual features alongside to the output of the LR model which was employed in the first layer.

We computed the statistical significance of the features extracted from the EMA questionnaires and passive sensing by employing the ANOVA test and Welch's t-test, respectively. The results of ANOVA test highlight the significant impact of activities, locations, and days of the week on perceived stress levels. In contrast, the results of the Welch's t-test showed that less than 20% of the features extracted from passive sensing are correlated to perceived stress level. Most of significant features are those related to activity features. However, intriguingly, features related to device usage specifically those related to screen state are not significant at all.

Although the Welch's t-tests and one-way ANOVA tests show some degree of contextual feature significance, aligning with the previous work [22], the results of LOSO CV technique showed only marginal improvement in accuracy upon the addition of EMA and passive sensing-based contextual features to physiological features. For future research, this may suggest that some degree of personalization is needed when using rich contextual information to further enhance accuracy.

B. Limitations and future directions

The findings from both laboratory and real-life assessments suggest that SOSW is a robust and practical methodology for stress detection. Its effectiveness in capturing stress-related changes, coupled with its integration with everyday smartwatches and contextual information, positions it as a valuable tool for monitoring and managing stress in various settings, from controlled environments to real-life situations. However, we would like to acknowledge that there exist certain factors that may impose limitations on its performance.

A factor that may impose limitation on the daily-life stress detection performance is the wide range of possible daily life scenarios that can potentially impact the user's perceived stress. While our methodology can accurately detect physiological stress arousals, there can be such instances where perceived stress exhibits less pronounced physiological arousal in real-life conditions, which may compromise the performance of our methodology. To mitigate such undesirable outcome, *SOSW* leverages rich contextual data encompassing behavioral biomarkers of perceived stress alonside the physiological arousal information.

Another factor that can limit the performance of stress detection methodology is the scale of a data collection study. With the number of participants involved our study, unfortunately, we may not be able to claim generalizability. A larger-scale and more longitudinal data collection conducted across diverse participant groups may be necessary to assert the generalizability of the stress detection methodology. This can cover a broader demographic spectrum, such as age, gender, and health backgrounds, and longer study durations (e.g., months or years) taking into account that individuals can adapt to certain scenarios and perception of stress can possibly vary over long periods of time. Last, but not least, a larger dataset would also enable effective utilization of advanced deep learning techniques such as deep neural networks for potentially improving stress detection accuracies.

The daily-life stress detection performance can also be limited by the battery life of the commercial smartwatches.

Although the battery life of the smartwatch Galaxy 5 guarantees more than 60 hours under typical use [69], continuous sensing may deteriorate the life span, which may limit the temporal coverage for physiological data acquisition. Another potential problem that may arise is the loss of data collection due to the interruption of data transmission via Bluetooth from the smartwatch to the smartphone. To ameliorate such a situation, the smartwatch can locally store the data and later, when the connection is reestablished, upload the sensing data.

It is also important to note that while the specific findings, presented in Section VI, are based on data collected from the smartwatch Galaxy 5, the underlying principles and methodologies can be applied to other devices with similar capabilities. Nevertheless, we acknowledge that variations in sensor accuracy and data processing algorithms across different devices may impact the exactness of the results. Therefore, as a future direction, we aim to expand the scope of our study to include a variety of devices, which will enhance the generalizability of our findings and provide a more comprehensive understanding of the capabilities and limitations of current smartwatch technology for stress detection.

VIII. Conclusion

In this study, we investigated the feasibility of using commercially available smartwatches combined with contextual data for detecting stress in both laboratory and real-life settings. To do so, we first conducted a preliminary study in which we analyzed the quality of physiological measurements obtained through the smartwatches. This early study revealed that physiological data collected through smartwatches are easily distorted by motion and loose wearing conditions. To cope with this, we proposed robust data processing pipelines. The *SOSW* methodology carefully combines motion artifact and looseness artifact removal techniques for improving measurements obtained through the smartwatches. Additionally, it considers a two-layer modeling architecture leveraging our best machine learning model obtained in our laboratory dataset for improving stress detection. To evaluate our methodology, we collected two datasets under laboratory and real-life conditions. For the laboratory dataset, using just physiological data obtained from off-the-shelf smartwatches, our proposed methodology can accurately detect stress instances with an F-1 score of 0.84. For the field dataset, our methodology can detect stress periods with an F-1 score of 0.66 using physiological data alone and an F-1 score of 0.71 by combining physiological and contextual data.

Acknowledgements

The authors would like to thank the editors and two anonymous reviewers for their constructive suggestions and helpful comments.

References

[1] B. S. McEwen, “Protection and damage from acute and chronic stress: allostasis and allostatic overload and relevance to the pathophysiology of psychiatric disorders,” *Ann. N. Y. Acad. Sci.*, vol. 1032, no. 1, pp. 1–7, 2004.

[2] Bracha and H. Stefan, “Freeze, flight, fight, fright, faint: Adaptationist perspectives on the acute stress response spectrum,” *CNS Spectr.*, vol. 9, no. 9, pp. 679–685, 2004.

[3] G. P. Chrousos, “Stress and disorders of the stress system,” *Nat. Rev. Endocrinol.*, vol. 5, no. 7, pp. 374–381, 2009.

[4] Y. S. Can, B. Arnrich, and C. Ersoy, “Stress detection in daily life scenarios using smart phones and wearable sensors: A survey,” *J. Biomed. Informatics*, vol. 92, pp. 103139–103161, 2019.

[5] R. Gordan, J. K. Gwathmey, and L. H. Xie, “Autonomic and endocrine control of cardiovascular function,” *World J. Cardiol.*, vol. 7, no. 4, pp. 204–216, 2015.

[6] W. Zhang, M. M. Hashemi, R. Kaldewaij, S. B. Koch, C. Beckmann, F. Klumpers, and K. Roelofs, “Acute stress alters the ‘default’brain processing,” *Neuroimage*, vol. 189, pp. 870–877, 2019.

[7] J. Park, “Work stress and job performance,” *Perspect. Labour Incom.*, vol. 8, no. 1, pp. 5–17, 2007.

[8] C. D. Conrad, A. M. Magariños, J. E. LeDoux, and B. S. McEwen, “Repeated restraint stress facilitates fear conditioning independently of causing hippocampal ca3 dendritic atrophy.,” *Behav. Neurosci.*, vol. 113, no. 5, pp. 902–914, 1999.

[9] V. Mishra, G. Pope, S. Lord, S. Lewia, B. Lowens, K. Caine, S. Sen, R. Halter, and D. Kotz, “Continuous detection of physiological stress with commodity hardware,” *ACM Trans. Comput. Healthcare*, vol. 1, no. 2, pp. 1–30, 2020.

[10] L. F. Barrett and A. B. Satpute, “Historical pitfalls and new directions in the neuroscience of emotion,” *Neurosci. Lett.*, vol. 693, pp. 9–18, 2019.

[11] K. Hovsepian, M. Al’Absi, E. Ertin, T. Kamarck, M. Nakajima, and S. Kumar, “cstress: towards a gold standard for continuous stress assessment in the mobile environment,” in *Proc. 2015 ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, pp. 493–504, 2015.

[12] Y. S. Can, N. Chalabianloo, D. Ekiz, J. Fernandez-Alvarez, C. Repetto, G. Riva, H. Iles-Smith, and C. Ersoy, “Real-life stress level monitoring using smart bands in the light of contextual information,” *IEEE Sensors J.*, vol. 20, no. 15, pp. 8721–8730, 2020.

[13] B. Egilmez, E. Poyraz, W. Zhou, G. Memik, P. Dinda, and N. Alshurafa, “Ustress: Understanding college student subjective stress using wrist-based passive sensing,” in *2017 IEEE Int. Conf. Pervasive Comput. Commun. Workshops*, pp. 673–678, IEEE, 2017.

[14] M. Gjoreski, M. Luštrek, M. Gams, and H. Gjoreski, “Monitoring stress with a wrist device using context,” *J. Biomed. Inform.*, vol. 73, pp. 159–170, 2017.

[15] T. Hao, K. N. Walter, M. J. Ball, H. Y. Chang, S. Sun, and X. Zhu, “Stresshacker: towards practical stress monitoring in the wild with smartwatches,” in *AMIA Annu. Symp. Proc.*, vol. 2017, pp. 830–839, American Medical Informatics Association, 2017.

[16] K. Kyriakou, B. Resch, G. Sagl, A. Petutschnig, C. Werner, D. Niederseer, M. Liedlgruber, F. Wilhelm, T. Osborne, and J. Pykett, “Detecting moments of stress from measurements of wearable physiological sensors,” *Sensors*, vol. 19, no. 17, pp. 3805–3831, 2019.

[17] M. Etiwy, Z. Akhrass, L. Gillinov, A. Alashi, R. Wang, G. Blackburn, S. M. Gillinov, D. Phelan, A. M. Gillinov, P. L. Houghtaling, and H. Javadikasgari, “Accuracy of wearable heart rate monitors in cardiac rehabilitation,” *Cardiovasc. Diagn. Ther.*, vol. 9, no. 3, pp. 262–272, 2019.

[18] H. K. Ra, J. Ahn, H. J. Yoon, D. Yoon, S. H. Son, and J. G. Ko, “I am a” smart” watch, smart enough to know the accuracy of my own heart rate sensor,” in *Proc. 18th Int. Workshop Mobile Comput. Syst. Appl.*, pp. 49–54, 2017.

[19] J. F. Horton, P. Stergiou, T. S. Fung, and L. Katz, “Comparison of polar m600 optical heart rate and ecg heart rate during exercise,” *Med. Sci. Sports Exerc.*, vol. 49, no. 12, pp. 2600–2607, 2017.

[20] D. K. Spierer, Z. Rosen, L. L. Litman, and K. Fujii, “Validation of photoplethysmography as a method to detect heart rate during rest and exercise,” *J. Med. Eng. Technol.*, vol. 39, no. 5, pp. 264–271, 2015.

[21] J. Ahn, H. K. Ra, H. J. Yoon, S. H. Son, and J. Ko, “On-device filter design for self-identifying inaccurate heart rate readings on wrist-worn ppg sensors,” *IEEE Access*, vol. 8, pp. 184774–184784, 2020.

[22] 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 *Proc. 2018 ACM Int. Joint Conf. Int. Symp. Pervasive Ubiquitous Comput. Wearable Comput.*, pp. 1708–1716, 2018.

[23] M. Nouman, S. Y. Khoo, M. P. Mahmud, and A. Z. Kouzani, "Recent advances in contactless sensing technologies for mental health monitoring," *IEEE Internet Things J.*, vol. 9, no. 1, pp. 274–297, 2021.

[24] 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, pp. 1849–1870, 2019.

[25] Y. S. Can, N. Chalabianloo, D. Ekiz, J. Fernandez-Alvarez, G. Riva, and C. Ersoy, "Personal stress-level clustering and decision-level smoothing to enhance the performance of ambulatory stress detection with smartwatches," *IEEE Access*, vol. 8, pp. 38146–38163, 2020.

[26] V. Mishra, S. Sen, G. Chen, T. Hao, J. Rogers, C. H. Chen, and D. Kotz, "Evaluating the reproducibility of physiological stress detection models," *Proc. ACM Interact. Mobile Wearable Ubiquitous Technol.*, vol. 4, no. 4, pp. 1–29, 2020.

[27] P. Siirtola, "Continuous stress detection using the sensors of commercial smartwatch," in *Adjunct Proc. 2019 ACM Int. Joint Conf. Pervasive Ubiquitous Comput. and Proc. 2019 ACM Int. Symp. Wearable Comput.*, pp. 1198–1201, 2019.

[28] R. Dai, C. Lu, L. Yun, E. Lenze, M. Avidan, and T. Kannampallil, "Comparing stress prediction models using smartwatch physiological signals and participant self-reports," *Comput. Methods Programs Biomed.*, vol. 208, pp. 106207–106218, 2021.

[29] K. Plarre, A. Raij, S. M. Hossain, A. A. Ali, M. Nakajima, M. Al'Absi, E. Ertin, T. Kamarck, S. Kumar, M. Scott, and D. Siewiorek, "Continuous inference of psychological stress from sensory measurements collected in the natural environment," in *Proc. 10th ACM/IEEE Int. Conf. Inf. Process. Sensor Networks*, pp. 97–108, IEEE, 2011.

[30] F. de Arriba Perez, J. M. Santos-Gago, M. Caeiro-Rodríguez, and M. J. F. Iglesias, "Evaluation of commercial-off-the-shelf wrist wearables to estimate stress on students," *J. Visualized Exp.*, no. 136, pp. e57590–e57599, 2018.

[31] J. Tervonen, S. Puttonen, M. J. Sillanpää, L. Hopsu, Z. Homorodi, J. Keränen, J. Pajukanta, A. Tolonen, A. Lämsä, and J. Mäntyjärvi, "Personalized mental stress detection with self-organizing map: From laboratory to the field," *Comput. Biol. Med.*, vol. 124, pp. 103935–103944, 2020.

[32] "Stress Management, Fitbit Technology." https://www.fitbit.com/global/us/technology/stress. Accessed on: Jan 15, 2024.

[33] "Measure your stress level with Samsung Health." https://www.samsung.com/us/support/answer/ANS00080574/. Accessed on: Jan 15, 2024.

[34] "Stress Tracking, Garmin Technology." https://www.garmin.com/en-US/garmin-technology/health-science/stress-tracking/. Accessed on: Jan 15, 2024.

[35] "The Health app, Apple." https://www.apple.com/ios/health/. Accessed on: Jan 16, 2024.

[36] "Smartwatch market share worldwide by vendor 2022 — Statista." https://www.statista.com/statistics/1296818/smartwatch-market-share/. Accessed on: Jan 15, 2024.

[37] B. D. Boudreaux, E. P. Hebert, D. B. Hollander, B. M. Williams, C. L. Cormier, M. R. Naquin, W. W. Gillan, E. E. Gusew, and R. R. Kraemer, "Validity of wearable activity monitors during cycling and resistance exercise," *Med. Sci. Sports Exerc.*, vol. 50, no. 3, pp. 624–633, 2018.

[38] R. K. Reddy, R. Pooni, D. P. Zaharieva, B. Senf, J. E. Youssef, E. Dassau, F. J. D. Iii, M. A. Clements, M. R. Rickels, S. R. Patton, and J. R. Castle, "Accuracy of wrist-worn activity monitors during common daily physical activities and types of structured exercise: evaluation study," *JMIR mHealth and uHealth*, vol. 6, no. 12, pp. e10338–e10356, 2018.

[39] N. Rashid, T. Mortlock, and M. A. A. Faruque, "Stress detection using context-aware sensor fusion from wearable devices," *IEEE Internet Things J.*, vol. 10, no. 16, pp. 14114–14127, 2023.

[40] H. Fukushima, H. Kawanaka, M. S. Bhuiyan, and K. Oguri, "Estimating heart rate using wrist-type photoplethysmography and acceleration sensor while running," in *2012 Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, pp. 2901–2904, IEEE, 2012.

[41] P. Renevey, R. Vetter, J. Krauss, P. Celka, and Y. Depeursinge, "Wrist-located pulse detection using ir signals, activity and nonlinear artifact cancellation," in *Proc. 23rd Ann. Int. Conf. IEEE Eng. Med. Biol. Soc. 2001.*, vol. 3, pp. 3030–3033, IEEE, 2001.

[42] D. Biswas, N. Simões-Capela, C. V. Hoof, and N. V. Helleputte, "Heart rate estimation from wrist-worn photoplethysmography: A review," *IEEE Sensors J.*, vol. 19, no. 16, pp. 6560–6570, 2019.

[43] Pankaj, A. Kumar, A. Ashdhir, R. Komaragiri, and M. Kumar, "Analysis of photoplethysmogram signal to estimate heart rate during physical activity using fractional fourier transform–a sampling frequency independent and reference signal-less method," *Comput. Methods Programs Biomed.*, vol. 229, p. 107294, 2023.

[44] 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 Investig.*, vol. 15, no. 3, pp. 235–246, 2018.

[45] T. Pereira, P. R. Almeida, J. P. Cunha, and A. Aguiar, "Heart rate variability metrics for fine-grained stress level assessment," *Comput. Methods Programs Biomed.*, vol. 148, pp. 71–80, 2017.

[46] B. Bent, B. A. Goldstein, W. A. Kibbe, and J. P. Dunn, "Investigating sources of inaccuracy in wearable optical heart rate sensors," *npj Dig. Med.*, vol. 3, no. 1, pp. 18–27, 2020.

[47] "Galaxy Watch 5." https://www.samsung.com/us/watches/galaxy-watch5/. Accessed on: Oct 15, 2023.

[48] "Polar - affordable heart rate monitoring chest straps." https://www.polar.com/en/sensors/h10-heart-rate-sensor. Accessed on: Aug 29, 2023.

[49] 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," *Cardiovasc. Diagn. Ther.*, vol. 9, no. 4, p. 379, 2019.

[50] H. Tanaka, K. D. Monahan, and D. R. Seals, "Age-predicted maximal heart rate revisited," *J. Am. Coll. Cardiol.*, vol. 37, no. 1, pp. 153–156, 2001.

[51] C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata, "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median," *J. Exp. Soc. Psychol.*, vol. 49, no. 4, pp. 764–766, 2013.

[52] K. A. Sidek and I. Khalil, "Enhancement of low sampling frequency recordings for ecg biometric matching using interpolation," *Comput. Methods Programs Biomed.*, vol. 109, no. 1, pp. 13–25, 2013.

[53] "Heartpy - python heart rate analysis toolkit." https://python-heart-rate-analysis-toolkit.readthedocs.io. Accessed on: Aug 29, 2023.

[54] J. A. Healey and R. W. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Trans. Intell. Transp. Syst.*, vol. 6, no. 2, pp. 156–166, 2005.

[55] S. D. Kreibig, F. H. Wilhelm, W. T. Roth, and J. J. Gross, "Cardiovascular, electrodermal, and respiratory response patterns to fear-and sadness-inducing films," *Psychophysiology*, vol. 44, no. 5, pp. 787–806, 2007.

[56] M. Kusserow, O. Amft, and G. Tröster, "Monitoring stress arousal in the wild," *IEEE Pervasive Comput.*, vol. 12, no. 2, pp. 28–37, 2012.

[57] M. Esterman, B. J. Tamber-Rosenau, Y. C. Chiu, and S. Yantis, "Avoiding non-independence in fmri data analysis: leave one subject out," *Neuroimage*, vol. 50, no. 2, pp. 572–576, 2010.

[58] 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," *Proc. ACM Interact. Mobile Wearable Ubiquitous Technol.*, vol. 3, no. 3, pp. 1–22, 2019.

[59] M. M. Rahman, R. Bari, A. A. Ali, M. Sharmin, A. Raij, K. Hovsepian, S. M. Hossain, E. Ertin, A. Kennedy, D. H. Epstein, and K. L. Preston, "Are we there yet? feasibility of continuous stress assessment via wireless physiological sensors," in *Proc. ACM Bioinformatics, Comp. Biol., Health Inf. Conf. (5th)*, pp. 479–488, 2014.

[60] T. Stütz, T. Kowar, M. Kager, M. Tiefengrabner, M. Stuppner, J. Blechert, F. H. Wilhelm, and S. Ginzinger, "Smartphone based stress prediction," in *User Model. Adapt. Personalization: 23rd Int. Conf. UMAP 2015*, pp. 240–251, Springer, 2015.

[61] W. Choi, S. Park, D. Kim, Y. Lim, and U. Lee, "Multi-stage receptivity model for mobile just-in-time health intervention," *Proc. ACM Interact. Mobile Wearable Ubiquitous Technol.*, vol. 3, no. 2, pp. 1–26, 2019.

[62] "Samsung Privileged Health SDK, Samsung Developer." https://developer.samsung.com/health/privileged/overview.html. Accessed on: Jan 15, 2024.

[63] "Wear OS Share Surges on Samsung's Highest Quarterly Smartwatch Shipments." https://www.counterpointresearch.com/insights/wear-os-share-surges-samsungs-highest-quarterly-smartwatch-shipments-q3-2021/. Accessed on: Jan 15, 2024.

[64] "Firebase cloud messaging (FCM) - a cross-platform messaging solution." https://firebase.google.com/docs/cloud-messaging. Accessed on: Oct 4, 2023.

[65] K. Toshnazarov, H. Baazizi, N. Narziev, Y. Noh, and U. Lee, "Easytrack-orchestrating large-scale mobile user experimental studies (poster)," in *Proc. 17th Annual Int. Conf. Mobile Syst. Appl. Serv.*, pp. 576–577, 2019.

[66] "gRPC - a high performance, open source universal rpc framework." https://grpc.io/. Accessed on: Nov 14, 2023.

[67] T. Zhu, P. Watkinson, and D. A. Clifton, "Smartwatch data help detect covid-19," *Nat. Biomed. Eng.*, vol. 4, no. 12, pp. 1125–1127, 2020.

[68] M. Fiordelli, N. Diviani, and P. J. Schulz, "Mapping mhealth research: a decade of evolution," *J. Med. Internet Res.*, vol. 15, no. 5, pp. e95–e109, 2013.

[69] "Galaxy watch5 specifications." https://www.samsung.com/uk/mobile-phone-buying-guide/galaxy-watch-5-vs-galaxy-watch-active-2-fit-2/. Accessed on: Jan 11. 2024.

Toshnazarov Kobiljon received the B.S. degree in Computer Science and Engineering from Inha University in Tashkent (IUT), Uzbekistan, in 2018, the M.S. degree in Electrical and Computer Engineering from Inha University, South Korea, in 2020. Currently, he is pursuing Ph.D. degree in Energy Engineering at Korea Institute of Energy Technology (KENTECH), South Korea. His research interests include human-computer interaction, and mobile and wearable-based mHealth application development, and cloud computing.

Uichin Lee received the B.S. degree in computer engineering from Chonbuk National University, in 2001, the M.S. degree in computer science from the Korea Advanced Institute of Science and Technology (KAIST), in 2003, and the Ph.D. degree in computer science from UCLA, in 2008. He continued his studies at UCLA as a Postdoctoral Research Scientist (2008–2009) and then worked for Alcatel-Lucent Bell Labs as a Member of Technical Staff (till 2010). He is currently an Associate Professor with the School of Computing, KAIST. His research interests include human–computer interaction (HCI), social computing, and ubiquitous computing.

Byung Hyung Kim received the master's degree in computer science from Boston University, Boston, MA, USA, in 2010, and the Ph.D. degree in computer science from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea, in 2018. He is an Assistant Professor with the Department of Artificial Intelligence, Inha University, Incheon, Republic of Korea. Before he joined Inha University, he was a Research Assistant Professor with the School of Computing, KAIST. His research interests include algorithmic transparency, interpretability in affective intelligence, computational emotional dynamics, cerebral asymmetry and the effects of emotion on brain structure for affective computing, brain–computer interface, and assistive and rehabilitative technology.

Varun Mishra completed their B.Tech. in Computer Science and Engineering at Shiv Nadar University, India, in 2015, and later earned a Ph.D. in Computer Science from Dartmouth College in 2021, focusing on digital interventions for mental and behavioral health. Currently, they are an Assistant Professor at Northeastern University, with roles in both the Khoury College of Computer Sciences and the Bouvé College of Health Sciences. Their research interests include novel sensing and intervention systems for smartphones and wearable devices, particularly in mental and behavioral health.

Lismer Andres Caceres Najarro received the B.Sc. degree from the Peruvian University of Applied Sciences, Lima, Peru, in 2010, the M.S. degree from Kyungsung University, Busan, Republic of Korea, in 2016, and the Ph.D. degree in electrical engineering and computer science from the Gwangju Institute of Science and Technology (GIST), Gwangju, Republic of Korea, in 2021. He is currently a research professor at Korea Institute of Energy Technology (KENTECH), South Korea. His research interests include target localization in wireless sensor networks, evolutionary algorithms, machine learning, multiagent systems, smart grids, and smart healthcare. Dr. Caceres Najarro was awarded the Graña y Montero Peruvian Engineering Research Award (fourth edition).

Youngtae Noh is an associate professor in the Department of Data Science, Hanyang University. He received his B.S. in Computer Science from Chosun University in 2005, an M.S. degree in Information and Communication from Gwangju Institute of Science Technology (GIST) in 2007, and a Ph.D. in computer science at the University of California, Los Angeles (UCLA) in 2012. Before joining Hanyang University, he worked as an associate professor at Inha University (2015 ∼ 2022), followed by KENTECH (2022 ∼ 2024). Moreover, he has also worked at Cisco Systems as a staff member (2012 ∼ 2014). His research areas include mobile/pervasive computing, mobile systems, mobile data science, mobile-HCI, cloud computing, data center networking, wireless networking, and future Internet.