Data Collection and Labeling of Real-Time IoT-Enabled Bio-Signals in Everyday Settings for Mental Health Improvement

Ali Tazarv
Electrical Eng. & Computer Science
University of California
Irvine, California, USA
atazarv@uci.edu

Sina Labbaf
Computer Science
University of California
Irvine, California, USA
slabbaf@uci.edu

Amir M. Rahmani
School of Nursing, Computer Science
University of California
Irvine, California, USA
a.rahmani@uci.edu

Nikil Dutt
Computer Science
University of California
Irvine, California, USA
dutt@uci.edu

Marco Levorato
Computer Sciecne
University of California
Irvine, California, USA
levorato@uci.edu

ABSTRACT

Real-time physiological data collection and analysis play a central role in modern well-being applications. Personalized classifiers and detectors have been shown to outperform general classifiers in many contexts. However, building effective personalized classifiers in everyday settings - as opposed to controlled settings - necessitates the online collection of a labeled dataset by interacting with the user. This need leads to several challenges, ranging from building an effective system for the collection of the signals and labels, to developing strategies to interact with the user and building a dataset that represents the many user contexts that occur in daily life. Based on a stress detection use case, this paper (1) builds a system for the real-time collection and analysis of photoplethysmogram, acceleration, gyroscope, and gravity data from a wearable sensor, as well as self-reported stress labels based on Ecological Momentary Assessment (EMA), and (2) collects and analyzes a dataset to extract statistics of users' response to queries and the quality of the collected signals as a function of the context, here defined as the user's activity and the time of the day.

CCS CONCEPTS

• **Computer systems organization → Sensor networks**; • **Human-centered computing → User studies**.

KEYWORDS

PPG in everyday settings, Health systems user behaviour, Health data labeling.

ACM Reference Format:
Ali Tazarv, Sina Labbaf, Amir M. Rahmani, Nikil Dutt, and Marco Levorato. 2021. Data Collection and Labeling of Real-Time IoT-Enabled Bio-Signals in Everyday Settings for Mental Health Improvement. In *Conference on Information Technology for Social Good (GoodIT '21), September 9–11, 2021, Roma, Italy.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3462203.3475918

1 INTRODUCTION

This study stems from the UNITE project housed at the University of California, Irvine. The goal of the project is to improve the well-being of pregnant women in underrepresented communities by integrating in-home visitations with wearable-based fine grain monitoring and interventions. The focus is on detecting and mitigating stress, a key indicator of pregnancy outcomes.

Within this context, in this paper we analyze the feasibility of an online learning strategy, whose final objective is building a stress detector by collecting a labeled dataset associating physiological signals to stress labels. Different from prior studies, we focus on every day settings, where there are no restrictions on movements and the environment, and subjects are doing normal daily routines.

In addition to the development of an effective system for the real-time collection of data, this scenario presents several inherent challenges related to the data toward the training of effective classifiers. Intuitively, one of the key challenges is the quality of the physiological signals, which may depend on several factors, including motion, and thus activity [2, 5]. However, a critical aspect is the response of the user to the queries (EMA). The user context may influence willingness to respond and the response time, thus affecting how representative the dataset is of the user's activities and the correlation between samples and labels.

This paper makes the following contributions:

- We build and deploy a three-tiered system for the collection and real-time analysis of labeled stress data. The system is composed of wearable sensors, an edge layer and a cloud server. We discuss system-level challenges that impact data acquisition capabilities.
- While the tiered system is capable of acquiring a large number of physiological signal samples, the user may be willing to label only a small fraction of them. We then develop a strategy to request labels based on the signal itself even in the absence of a prebuilt classifier. The strategy aims to collect a number of samples proportional to the density of samples in the feature space, but also capture outliers and rare events.

GoodIT '21, September 9–11, 2021, Roma, Italy
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8478-0/21/09…$15.00
https://doi.org/10.1145/3462203.3475918

- We collect a dataset from a group of volunteers in everyday settings. Specifically, we collect biosignals photoplethysmogram (PPG) along with the movement data – Acceleration, Gyroscope and Gravity – from sensors on a smart watch in a real-time scheme, and collect self reported stress levels from participants. The raw signals (PPG, ACC, Gyro and Gra) are collected in a window of 2 minutes, once every 15 minutes. The list of labels we collect includes mental stress level, emotional status and physical activity.
- We study the quality of the signals and the willingness of the user to respond to queries as a function of the current context, here defined in terms of user's activity and the time of the day. We also show distribution plots of response time and response rate of users, during our experiment. On average the response rate is usually lower in early morning (7-9 AM) and higher in early afternoon (2-3 PM), and response time shows a different distribution for different activities. We analyze the quality of PPG signals for each predefined activity separately, and show that during the activities with less movements, the measured signal has better quality. We analyze the temporal correlation of samples as a function of context (activities) and show that consecutive samples are more similar to each other, compared to distant samples.

The rest of the paper is organized as follows: This paper starts by describing the system model and architecture used for data collection and sending queries for labels, highlighting the challenges introduced by the system constraints in Section 2. Section 3 describes our strategy for sending queries and collecting labels for the data optimally in the absence of a prebuilt classifier. Section 4 describes our approach and the methods we used for analysis of the collected data. Section 5 presents the results of the analysis, in terms of coverage of sample space with labeled data, contextual analysis, signal quality, and user behaviour analysis in different contexts and on different self reported stress levels. Finally, Section 6 concludes the paper with a summary and directions for future work.

2 SYSTEM MODEL

First, we describe the system we developed to enable data collection and real-time interaction with the user. As shown in Fig. 1, the system is composed of three tiers: sensor layer, edge layer and cloud layer. The sensor layer which is a smart watch, collects the raw signals while the cloud layer performs feature extraction and other computationally expensive and power consuming tasks, including selecting a portion of the samples to be labeled by the user.

We provide users with an interface to report labels in the form of a smart-phone application. The smartphone app asks the participants to label the samples through an Ecological Momentary Assessment (EMA), in which push notifications queries the participant about their stress level, recent physical activity or physical state (e.g. sitting, standing, etc.). The phone also functions as a gateway, building a connection path for the smart watch to transfer the sensor data to the cloud layer (through the internet connection on the phone). In the following, we describe each layer in detail and discuss the challenges we encountered for data collection using wearables in everyday settings.

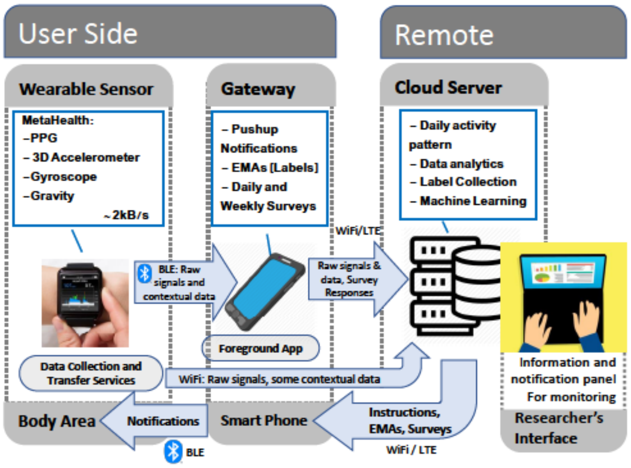

Figure 1: Overview of the system architecture.

2.1 Sensor Layer

We selected a wearable platform capable of acquiring and transmitting raw physiological (PPG) and motion (Accelerometer, Gyroscope and Gravity) signals. Specifically, we use Samsung Gear Sport smartwatches and developed a service in Tizen that is capable of collecting raw PPG, accelerometer, gyroscope and gravity signals and sending them to the cloud in real-time. The sampling frequency of the sensors is $20Hz$. The watch can send the data directly to the cloud layer if it is connected to a local Wi-Fi (path 1) or through through the internet connection on the smartphone (path I) as explained later.

The data collection application on the sensor layer includes two services and a user interface (UI). The first service collects the sensor data at a constant rate (once every 15 minutes) and duration (2-minute intervals) and sends it to the cloud in real-time. If the service fails to send the data immediately, the data sample is stored on the watch and transferred to the server at a later time. The UI is a simple app on the watch for restarting these two services.

2.2 Cloud Layer

A cloud web-server receives the data samples from the watch and immediately initiates processing. Based on the observed features of each incoming sample, an internal logic determines whether or not to trigger the EMA to collect a label from the user. If triggered, it sends a signal to the smartphone and a push notification appears on the screen, if the user opens it, the UI (a phone app) shows a questionnaire composed of simple questions corresponding to labels. The responses are then transferred to the cloud. The samples, features and labels are stored in a MongoDB database.

2.3 Edge Layer

The sensor is connected to a smartphone via Bluetooth Low Energy (BLE). If the watch is not connected to a local Wi-Fi, in order to send the collected data to the cloud, the watch proxies the phone's internet connection (path 2) through BLE. This setting is energy efficient, and thus suitable for everyday setting applications. This back up connection route is designed to take effect when the watch is not directly connected to a local Wi-Fi router. Additionally, we designed a UI for the smartphone (android and iOS apps) in order to communicate with the users and collect the labels.

2.4 Challenges

From a system perspective, the first design challenge is to set the monitoring duration and frequency such that: 1. The total delay from data acquisition to the EMA notification is tolerable, 2. Power consumption matches the characteristics of the devices and user requirements, 3. The sample quality allows meaningful feature extraction, and 4. The signals provide a meaningful representation of temporal variations. Note that the capacity of the wireless channel connecting the watch to the smartphone is extremely limited.

2.4.1 End-to-End Delay. First, we measured the bandwidth of the BLE proxy Internet connection manually to characterize hardware limitations. The maximum data-rate is approximately $80kB/s$. As the sensor generates $2KB/s$ of data, the upload time for each minute of data is only 1.5s. The processing time at the server is negligible compared to the upload time.

After collecting and sending the EMA, there is a small delay until the notification appears on the phone. This delay depends on internet connection, type of the phone and the mode the phone is operating on (power saving, etc.). However, this delay is typically small compared to the users' response time.

2.4.2 Power Limitations on the Watch. To measure power consumption, we first used the watch without any monitoring services and measured the battery lifetime, and compared it with that observed with the monitoring system continuously active. The measured battery lifetimes are $\sim$ 40 hours and $\sim$ 10 hours, respectively. In order to extend battery lifetime to 24 hours (to allow for nightime recharging), we then need to keep the monitoring system active for at most 22% of the time. This limits our ability to collect continuous signals over extended periods of time, and raises the issue of how to shape a parsimonious sample collection strategy.

2.4.3 Sampling Times and Signals Duration. We determine sample collection based on the considerations above. Importantly, the data collection window plays an important role in the quality of PPG signals. However, if the window duration is increased, then we are forced to do the measurement fewer times throughout the day. As shown in [1], a 2-minute time window of PPG/ECG signal can provide us with sufficiently accurate extraction of the majority of Heart Rate Variability (HRV)[1] features. Hence, By setting the minimum duration of data samples to 2 minutes, sampling every 15 minutes will satisfy all system constraints described above while extending the battery life to up to 42 hours in practice.

3 DATA COLLECTION

The ultimate goal of data collection is to train personalized classifiers that can detect mental health status based on biosignals (PPG) and signals describing motion (Accelerometer, Gyroscope and Gravity). One of the key challenges in collecting such datasets in everyday settings is the interaction with the users, as sending queries for labeling too often can be overwhelming and may lower response rate and eventually degrade the dataset. Intuitively, the system should parsimoniously trigger the EMA to collect a label to build a meaningful dataset as quickly as possible without imposing excessive burden on the user. To this aim, we devised a selection method that triggers the labeling query based on statistics of the previously collected and current samples.

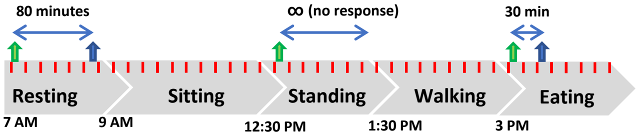

Figure 2: Example of Data collection in everyday setting.

Figure 2 represents such a data collection scenario. Samples are collected once every fifteen minutes (red pointers), green arrows correspond to the EMA notifications, and the vertical blue arrows correspond to responses from the user.

The list of labels we collect through the EMA includes stress levels (*not at all*, *a little bit*, *some*, *a lot*, and *extremely*), emotions (*sad, mad, neutral or happy*), and recent activities or physical status (e.g., sitting, walking, jogging, etc.). Later in this section we describe how we select a portion of samples to be labeled.

3.1 Data Cleaning and Feature Extraction

Before we apply the selection method, we pre-process the raw signals and extract the corresponding features (using HeartPy Package [6]). When the raw PPG sample is received at the cloud layer, it is first passed through a Butterworth band-pass filter to clean up the high and low frequency noises. The band-pass filter is of order 3, with cut off frequencies set at $(0.7Hz, 3.5Hz)$, corresponding to $42bpm$ and $210bpm$) respectively. Then the signal is passed through a moving average filter (length = 0.75 seconds). We, then, apply a peak detector to the filtered signal. Using the peak points of the filtered signal, we extract thirteen features from each sample (window). These features are: BPM, IBI, SDNN, SDSD, RMSSD, pNN20, pNN50, MAD, SD1, SD2, S, SD1/SD2, and BR[2]. We use these features for further processing and decision makings.

3.2 Strategy for Labeling Selected Data

Data collection consists of an Initial Phase and a Query Phase:

Initial Phase: We start the procedure by observing the first N samples for each subject to get an estimate of the distribution of samples in the sample space for that subject. In this phase we do not collect any labels. We used N = 100 in our experiment. It is important to note that different subjects might have different patterns in their PPG which can result in personalized distributions in the sample space. This motivates the need to estimate the distribution of the data for each subject separately.

Query Phase: For samples after the initial phase, we trigger the EMA for a subset of samples. The probability of triggering the EMA

[1] Heart Rate Variability or HRV is a set of informative features including statistics of heart beats that can be extracted from PPG signals.

[2] **BPM:** Beats per Minute, Heart Rate. **IBI:** Inter-Beat Interval, average time interval between two successive heart beats (called NN intervals). **SDNN:** Standard Deviation of NN intervals. **SDSD:** Standard Deviation of Successive Differences between adjacent NNs. **RMSD:** Root Mean Square of Successive Differences between the adjacent NNs. **pNN20:** The proportion of successive NNs greater than 20ms (or 50ms for pNN50). **MAD:** Median Absolute Deviation of NN intervals. **SD1 and SD2:** Standard Deviations of the corresponding Poincaré plot. **S:** Area of ellipse described by SD1 and SD2. **BR:** Breathing Rate.

for each sample is proportional to the density of the region of the sample. This way, if a sample falls in a region in which there has been a large number of unlabeled samples it is more likely that we trigger the EMA. For each region, after we collect sufficient number of labels, we stop collecting labels. However, all the probability values (for all the regions) are *clipped* on the bottom at P = 0.1. So if a samples falls in a region where there is little or no previous samples, the probability of query is still non-zero. This results in exploring unseen regions, as well as more dense regions.

The ultimate goal of this experiment is to train a personalized classifier to predict mental health conditions. In order to do that, we need to collect enough labeled data, such that they optimally cover the sample space. To measure how the labeled samples cover the sample space we need a measure.

If we define all the incoming samples as $X = \{x_0, x_1, ..., x_M\}$, labeled samples as $U_i = \{u_0, u_1, ..., u_i\}$ ($U_i \subset X$) and the corresponding labels as $Y_i = \{y_0, y_1, ..., y_i\}$ (Y_i is the set of labels collected up to label number i), we then define the coverage metric as:

$$F_D(i) = \frac{\mathbf{card}(\ \{x \in X \mid \|x - u^*\| > D\}\ )}{\mathbf{card}(\mathbf{X})}$$

in which u^* is the closest labeled sample to the sample x, **card** is the cardinality of the set, and D is a distance constant. Note that u_k or x_k are each a vector of 13 elements (features), extracted from a PPG sample (window size 2 minutes). As we collect more labels, we count the number of samples which are farther than a certain threshold D to the closest labeled sample collected up to that point. The ratio between this number and the total number of samples gives us a metric $F_D(i) \in [0, 1]$. The smaller this metric, the better labeled samples represent the entire data.

4 PROCESSING AND EVALUATION METHODS

4.1 Signal Quality

The quality of PPG signal is among the most important factors to consider when collecting data in everyday settings. Due to the architecture of PPG sensors, PPG signals are highly prone to motion and noise artifacts (MNA) [3], which can make them unreliable, especially in everyday-settings. Therefore, a signal quality assessment on the collected signal is essential.

Several quality assessment indices for PPG signals are proposed in the literature. In this study we use five different indices [4]: Variation in Skewness of Heart Cycles measures the variations in asymmetry in distribution of peaks in the heart cycles; Variation in Kurtosis of Heart Cycles evaluates the variations in flatness or peakedness level of the heart cycles; Variation in approximate Entropy of Heart Cycles evaluates the variations in complexity of the heart cycles; Shannon Entropy obtains the level of noise in a segment of PPG signal; and Spectral Entropy calculates the signal complexity in frequency domain. For all of these five indices the lower values indicate a more *reliable* signal. These indices are extensively described in [4] with formal definition for each.

4.2 Activity Detection

We collect data from Accelerometer, Gyroscope, and Gravity sensors as well as the PPG signal. Each of these signals (except PPG) are measured in 3 dimensions (x, y, z) and can be used to detect user's activity during each measurement. The MNA in PPG signals show different patterns depending on users' activities. So if we partition samples based on the type of activity, data in each partition might be easier to analyze.

Table 1: Size of partitions based on predicted activity

ACTIVITY:	Sit	Stand	Walk	Jog	Others	Total
Samples:	31,040	5,919	3,849	12	33,254	74,074
Percent:	%41.9	%8.0	%5.2	%0.016	%44.9	%100

To do that, we need an activity detector. We used a publicly available dataset [7] and trained a Random Forest classifier on this dataset to build an activity detector. The external dataset consists of accelerometer and gyro data from 51 subjects, and labeled with 18 different activities. We selected the activities that are expected to be more common in everyday settings (*sitting, standing, walking, and Jogging*), kept these labels and changed the rest of them to *others.* We then trained our activity detector on this dataset. After fine tuning the model parameters, the activity detector showed 84% accuracy on leave-two-subjects-out evaluation method (trained on 49 subjects and tested on two subjects). We then used this trained classifier to predict the dominant activity during the measurement of each sample in our dataset. Then we partition our collected data into subsections based on those activities. The distribution of predicted activities on the entire dataset is shown in Table 1.

5 DATA ANALYSIS AND RESULTS

We proposed a labeling query engine that determines (in real-time) what samples are interesting to be labeled by the user. We collect raw biosignals (PPG) and raw movement data in durations of 2 minutes, up to 4 times per hour from fourteen volunteers (ten males and four females) over periods of time between 1 week to 3 months for different subjects. The corresponding labels come from self reported EMAs which are triggered for some samples. Table 2 presents the total number of samples we have collected throughout the experiment, along with the number of labels, and the number of labels that could be assigned to a sample (labels that had a sample withing 16 minutes around them).

We now analyze the dataset, using the methods and metrics defined in Section 4. Analysis include coverage of sample space, temporal correlations of samples, quality of PPG signals, and response time/rate of subjects in different contexts in everyday settings.

5.1 Sample space Coverage

In section 3 we defined a coverage metric $F_D(i)$ that quantifies how well the labeled samples cover the sample space. Labeled samples that cover the sample space proportionally and optimally are an important requirement for training a classifier that predicts human health and well-being. A plot of this metric over the number of labeled samples for one subject and for the threshold distance $D = 1.5$ is presented in Figure 3. In this plot the sample space coverage metric goes from around 1 to less than 0.1 as we collect the first 100 labels (generally it depends on the response rate but for this user it took about 3 weeks).

Table 2: Number of samples and labels for each subject

Subject	S01	S02	S03	S04	S05	S06	S07	S08	S09	S10	S11	S12	S13	S14	Total
Samples	4,580	2,164	1,764	2,580	2,267	17,552	10,087	2,752	1,236	7,910	2,555	12,296	3,738	1,332	74,074
Total labels	228	101	67	56	68	376	105	96	53	119	73	956	47	61	2,406
Used labels	217	92	42	53	59	370	101	93	50	104	60	942	45	55	2,283

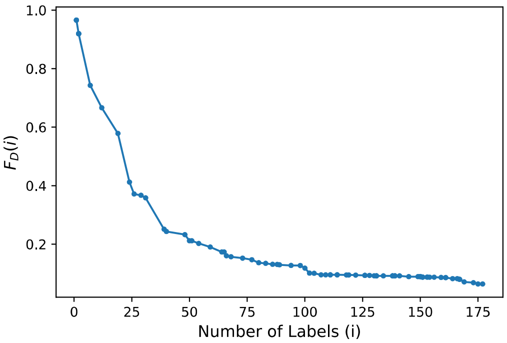

Figure 3: Fraction of unlabeled data that are farther than distance D = 1.5 from the closest labeled point.

5.2 Temporal correlation of samples

In the procedure of collecting labels, sometimes users cannot respond to the push notifications quickly enough for the sample to be meaningful. One important question then is how quickly humans' status (e.g., stress level, emotional status, and the corresponding physiological effects) changes. We submit queries for labels only a few times each day and obtain responses for a fraction of them. As a result, the correlation between the sample and the label may degrade. Thus, we perform a temporal analysis of the collected samples for several contexts. Specifically, we observe how the average distance of samples evolves in the sample space over time.

Two plots of the average distance of samples that are T minutes apart are presented in Figures 4 and 5 for various contexts and stress levels, on the data from one subject. The important takeaway from these two plots is that consecutive samples (15 minutes apart) are similar to one another, but for samples that are farther than 15 minutes apart, there is no significant or consistent similarity (the average distance almost saturates at 30 minutes and after). This pattern is consistent among various contexts, various stress levels, and also various subjects.

In summary, we observe that different activities and stress levels influence the coherence of the collected samples, which may affect the ability of the system to detect stress levels in various activities and user situations.

5.3 Data Quality Analysis

Quality of PPG signals is one of the biggest challenges whilst collecting data in everyday-settings, therefore a quality analysis of the collected data is essential. We use the five SQIs introduced in section 4.1 for the quality assessment of samples, and perform a separate analysis for each activity.

The distributions of the five SQI for various activities are shown in Figure 6. As expected, with increased motion all the five indices show that data quality decreases among the data marked sitting, standing, walking, respectively. Since the number of samples predicted as "Jogging" was too small, we didn't include that in this analysis. We used the data from all the fourteen subjects.

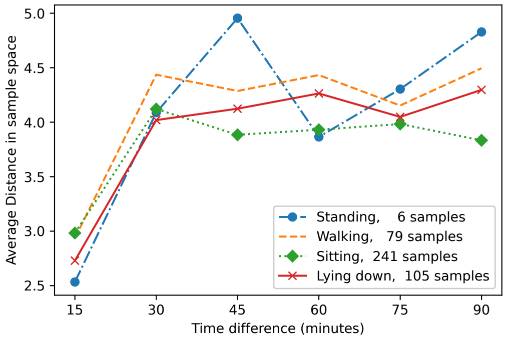

Figure 4: Average distance of samples (after normalizing the features) VS. their time distance for subject S12.

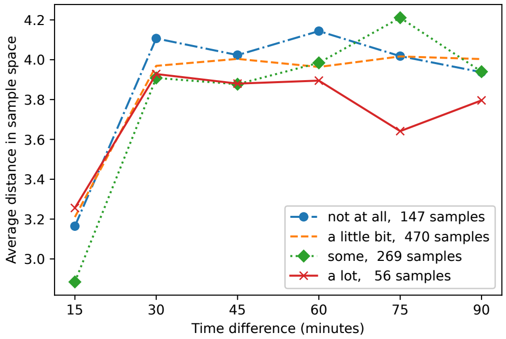

Figure 5: Average distance of samples VS. their time difference, for subject S12, for different reported stress levels.

5.4 Response Rate and Response time

Response time is defined as the time it takes a user to respond an EMA after they receive the notification. In Figure 7 we present the cumulative density of response time while being in certain contexts (e.g. sitting, etc.). We can see from these CDF plots that users are less likely to respond to the EMA faster when they are *Lying down*, compared to other physical states (or activities).

Similar to the different activities, response time shows different patterns when users are in different reported stress levels. Figure 8 presents the CDF and the median (where the CDF is 0.5) for different self reported stress levels. The pattern suggests that users are more likely to respond faster when they are in stressful situations.

Response rate, defined as the ratio of the number of responses to the number of queries over a period of time, is also analyzed

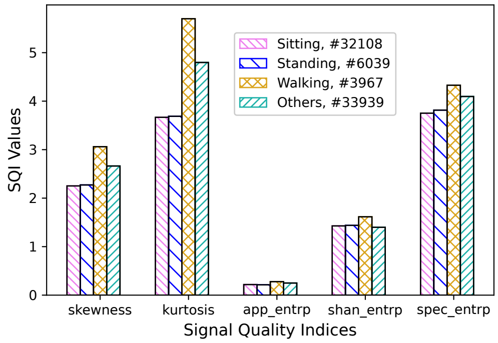

Figure 6: Signal Quality Indices for various predicted activities. A lower index means the sample is more reliable.

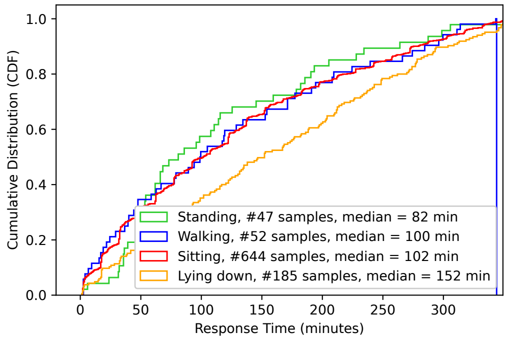

Figure 7: Probability of getting a response within a time frame while doing various self reported activities.

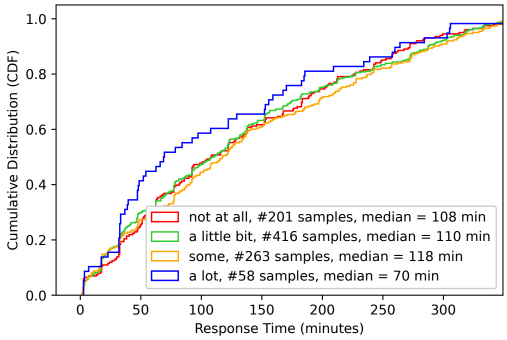

Figure 8: Probability of getting a response within a timeframe while in various self reported mental states.

here. Response rate varies at different hours of day, and also follows different patterns among different subjects, but still some common patterns can be observed among all subjects. Figure 9 shows the average response rate over all the subjects for different hours of day; response rate tends to be lower in early morning and higher in early afternoon.

6 CONCLUSIONS AND FUTURE WORK

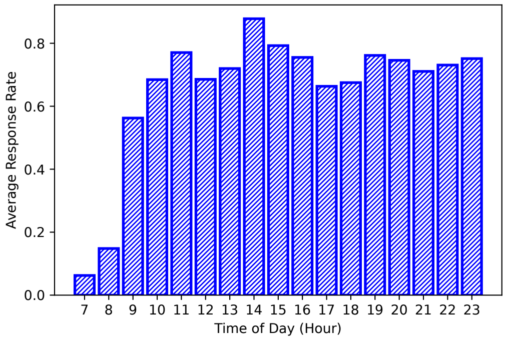

Figure 9: Average response rate vs. time of day

Collecting photoplethysmogram (PPG) signals with corresponding self reported labels in everyday settings is a big challenge. Our study used the Samsung Gear Sport smart-watch as a sensor device in a system for this phase of data collection and utilized a method to improve the label collection procedure. The data were collected from fourteen active volunteers in everyday settings. We tested our personalized label query engine and performed a set of contextual analysis on the data. The quality analysis on the biosignals confirms that these signals are more reliable in certain predictable contexts (i.e. in lower physical movements). A temporal analysis of the data shows that consecutive samples for one subject are similar, but there is no consistent similarity between samples that are more than 15 minutes apart. In addition, an analysis on users behaviour shows some clear patterns in response times and response rates at different hours of day, and under different mental and physical status.

These observations motivate our future work that will utilize more sophisticated methods (possibly variants of active learning) in the labeling process. Better labels will allow us to design a classifier that can possibly detect mental health conditions of the users based on biosignals and contextual information – promising to provide valuable tools for mental health professionals to better diagnose and treat emotional and stress problems in a context-aware and personalized manner.

ACKNOWLEDGMENTS

This work was partially supported by NSF Smart and Connected Communities (S&CC) grant CNS-1831918.

REFERENCES

[1] Hyun Baek, Chul-Ho Cho, Jaegeol Cho, and Jong-Min Woo. 2015. Reliability of Ultra-Short-Term Analysis as a Surrogate of Standard 5-Min Analysis of Heart Rate Variability. *Telemedicine journal and e-health : the official journal of the American Telemedicine Association* 21 (2015). https://doi.org/10.1089/tmj.2014.0104

[2] Hee Jeong Han et al. 2020. Objective stress monitoring based on wearable sensors in everyday settings. *Journal of Medical Engineering & Technology* 44, 4 (2020), 177–189.

[3] Yuka Maeda, Masaki Sekine, and Toshiyo Tamura. 2011. Relationship between measurement site and motion artifacts in wearable reflected photoplethysmography. *Journal of medical systems* 35, 5 (2011), 969–976.

[4] Aysan Mahmoudzadeh, Iman Azimi, Amir M Rahmani, and Pasi Liljeberg. 2021. Lightweight Photoplethysmography Quality Assessment for Real-time IoT-based Health Monitoring using Unsupervised Anomaly Detection. *Procedia Computer Science* 184 (2021), 140–147.

[5] Emad Kasaeyan Naeini et al. 2019. A Real-time PPG Quality Assessment Approach for Healthcare Internet-of-Things. *Procedia Computer Science* 151 (2019).

[6] Paul van Gent, Haneen Farah, N Nes, and Bart van Arem. 2018. Heart rate analysis for human factors: Development and validation of an open source toolkit for noisy naturalistic heart rate data. In *Proceedings of the 6th HUMANIST Conference*. 173–178.

[7] Gary M Weiss. 2019. WISDM Smartphone and Smartwatch Activity and Biometrics Dataset. *UCI Machine Learning Repository: WISDM Smartphone and Smartwatch Activity and Biometrics Dataset Data Set* (2019).