2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Personalized Stress Monitoring using Wearable Sensors in Everyday Settings

Ali Tazarv‖, Sina Labbaf*, Stephanie M. Reich†, Nikil Dutt*,§, Amir M. Rahmani‡,*,§, Marco Levorato*
‖ Dept. of Electrical Engineering and Computer Science, ∗ Dept. of Computer Science,
† School of Education, ‡ School of Nursing, § Institute for Future Health (IFH)
University of California, Irvine

***Abstract*— Since stress contributes to a broad range of mental and physical health problems, the objective assessment of stress is essential for behavioral and physiological studies. Although several studies have evaluated stress levels in controlled settings, objective stress assessment in everyday settings is still largely under-explored due to challenges arising from confounding contextual factors and limited adherence for self-reports. In this paper, we explore the objective prediction of stress levels in everyday settings based on heart rate (HR) and heart rate variability (HRV) captured via low-cost and easy-to-wear photoplethysmography (PPG) sensors that are widely available on newer smart wearable devices. We present a layered system architecture for personalized stress monitoring that supports a tunable collection of data samples for labeling, and present a method for selecting informative samples from the stream of real-time data for labeling. We captured the stress levels of fourteen volunteers through self-reported questionnaires over periods of between 1-3 months, and explored binary stress detection based on HR and HRV using Machine Learning methods. We observe promising preliminary results given that the dataset is collected in the challenging environments of everyday settings. The binary stress detector is fairly accurate and can detect stressful vs non-stressful samples with a macro-F1 score of up to %76. Our study lays the groundwork for more sophisticated labeling strategies that generate context-aware, personalized models that will empower health professionals to provide personalized interventions.**

I. Introduction

Stress can contribute to illness through its direct physiological effects or indirectly through maladaptive health behaviors (*e.g.*, smoking, poor eating or sleeping habits) [1]. It is therefore critical to motivate people to adjust their behavior and lifestyle and introduce appropriate strategies to achieve a better stress balance before an increased level of stress results in serious health conditions [2].

The increasing availability of wearables, interconnected devices capable of acquiring high-quality biosignals, opens important opportunities for advanced machine learning-enabled health monitoring and intervention applications [3], [4]. Recent literature [5] demonstrates that it is indeed possible to objectively detect stress by analyzing biological signals. However, existing objective stress detection frameworks are designed for controlled settings, where data is recorded while users are in a set of predefined physical states or performing certain activities (*e.g.* sitting, lying down, running). On the other hand, stress detection needs to be performed in *everyday settings*, where subjects are engaged in their normal daily activities and routines. Everyday settings pose inherent challenges for stress monitoring, including: real-time collection and analysis of data; the lower quality of signals due to motion and noise artifacts (MNAs); and difficulties in collecting self-reports due to limited user adherence [5], [6]. Furthermore, personalization of stress monitoring in everyday settings raises additional challenges: features may emerge that are specific to the user's characteristics, behavioral patterns, physiology and context, as well as sensor setup/configuration, thus presenting a much higher degree of variations from one person to another compared to controlled settings. These differences can result in degrading the performance of general classifiers in everyday settings.

Since effective everyday stress monitoring and intervention must be personalized and context-aware, the underlying ML (Machine Learning) core needs to be adapted to match the stream of data generated by the user. It is important to note that personalized classifiers often outperform those trained using data from the general population even in idealized settings [5]. These classifiers depend on labels generated from subjects in-the-moment to accurately record instances of stress. However, users may not respond in a timely or interactive manner, resulting in a trade-off between the number of labels provided by the subjects versus the accuracy of the predictive model. This trade-off creates the need for a smart label query strategy that we use to explore real-time stress detection based on wearable data in every day settings.

The key contributions of this paper are:

- We architect a three-tier system for the collection and real-time analysis of biosignals labeled using self-reports. The system is composed of wearable sensors, a smartphone serving as a gateway, and a cloud server. We discuss system-level challenges influencing data acquisition capabilities.
- We develop a smart strategy to obtain labels for an adequate number of samples to proportionally represent the entire data from each user while capturing less overlapping regions of the feature space.
- We develop a machine learning based stress predictor. We map the stress labels into binary *stressed (1)* and *not stressed (0)*, and then train classification methods using these labels.
- We capture the stress levels of fourteen volunteers through self-reported questionnaires over periods of between 1-3 months, and evaluate our binary stress detection based on HR and HRV. Our classifiers are able to identify the binary

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

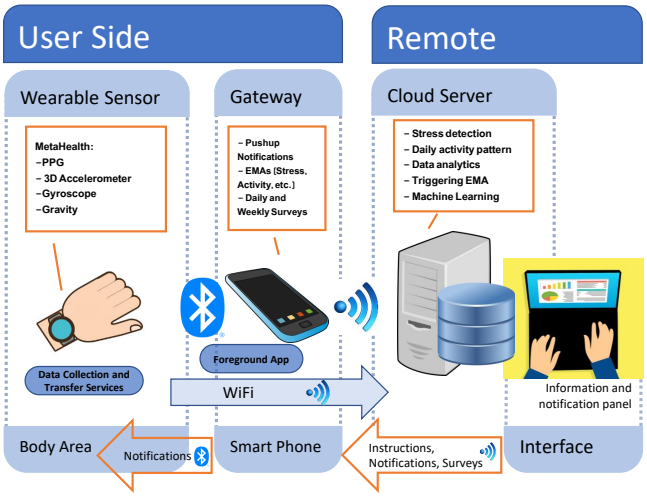

Fig. 1. Overview of the system architecture.

classes with an F1 score up to 76%. We also analyze the effect of personalization, and show how the stress detection performance improves over time, as we collect more labels from a subject and use those in the training process.

The rest of the paper is organized as follows. Section II describes the system architecture we used for data collection, and the proposed strategy for label collection in everyday settings. Section III describes the classification methods we used for detecting stress. Section IV presents our analysis and results on stress detection. Section V concludes the paper with a summary and directions for future work.

II. System Model and Data Collection

The ultimate goal of data collection is to train a personalized classifier for stress detection based on biosignals (PPG). 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 may overwhelm the users and may also lead to unnecessary data collection.

To enable real-time interaction with the user while minimizing resource usage, a layered design is necessary. The sensor layer should collect the raw signals, while the cloud layer processes the data, determines the quality of the signals, performs feature extraction and other computationally intensive and power consuming tasks. We also have to provide users with an interface to input their labels and build a communication path between these layers. In this section, we present our proposed solution for these requirements.

The experimental procedures involving human subjects described in this paper were approved by the Institutional Review Board (IRB) at UC Irvine.

A. System Architecture

Figure 1 outlines the system we developed to acquire the real-time dataset associating self-reported stress ratings from users to biosignals from wearable devices. We use a 3-tier architecture composed of a wearable device (sensor layer), a smartphone (edge layer), and a remote cloud server (cloud layer) working to collect and process the data, as described below.

*1) **Sensor Layer**:* The wearable platforms acquire and transmit raw physiological (PPG) and Movements (Accelerometer, Gyroscope and Gravity) signals. We used Samsung Gear Sport smartwatches and developed a service running on Tizen operating system to collect raw PPG and movement signals. The sampling frequency of all the above mentioned signals is $20Hz$. The watch can send the data directly to the cloud layer (if connected to a local Wi-Fi), or to the smartphone via Bluetooth. The raw signal acquisition application 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-minutes intervals) and sends it to the cloud. If the service fails to send the data to the cloud immediately, the data will be stored on the watch and transferred to the server at a later time. The UI is a simple app installed on the watch for restarting these two services.

*2) **Cloud Layer**:* A cloud web-server receives the data samples from the watch and processes them. Based on the observed features of each incoming sample, an internal logic (described later) determines whether it needs to ask users for a label or not. The responses from users are transferred to the cloud and stored in the database.

*3) **Edge Layer**:* We developed a smartphone app that asks the participants for labels through an Ecological Momentary Assessment (EMA). The EMA is triggered by the cloud for a portion of samples and when triggered, a push notification is displayed on the phone that asks the participant about their stress levels and recent physical activity or state (e.g., sitting, standing, etc.). The stress levels in the EMA are *not at all*, *a little bit*, *some*, *a lot*, and *extremely*.

The edge-cloud connection is established through the Internet on the smartphone. In addition, the watch is connected to the smartphone using Bluetooth Low Energy (BLE). In order to send the collected data to the cloud, the watch proxies the connected phone's Internet connection through BLE. This setting is energy efficient, and thus suitable for everyday setting applications. This is a back up connection and takes effect when the watch is not directly connected to the Wi-Fi.

B. Data Labeling

The system needs to parsimoniously trigger the EMA to collect labels to build a meaningful dataset as quickly as possible without imposing excessive burden on the user. Therefore, we devised a selection method to select informative samples to be labeled by the user. However, before applying the selection method, the raw signals need to be pre-processed for extracting the corresponding features.

1) Data Cleaning and Feature Extraction: When a raw PPG sample is received at the cloud, we first filter the raw signal to clean up the high and low frequency noises. We apply a Butterworth band-pass filter of order 3, with cut off frequencies set at $(0.7Hz, 3.5Hz)$, corresponding to $42bpm$ and $210bpm)$ respectively. Then, we pass the signal through a moving average filter and at the end apply a peak detector on it. Using the peak points of the filtered signal, we extract

TABLE I

NUMBER OF SAMPLES AND LABELS COLLECTED FROM EACH SUBJECT

Subject	# samples	total labels	used labels
S01	4,580	228	217
S02	2,164	101	92
S03	1,764	67	42
S04	2,580	56	53
S05	2,267	68	59
S06	17,552	376	370
S07	10,087	105	101
S08	2,752	96	93
S09	1,236	53	50
S10	7,910	119	104
S11	2,555	73	60
S12	1,2296	956	942
S13	3,738	47	45
S14	1,332	61	55

thirteen features from each sample (2 minutes of data). These features are: BPM, IBI, SDNN, SDSD, RMSSD, pNN20, pNN50, MAD, SD1, SD2, S, SD1/SD2, and BR[1]. We use these features for further processing and decision making.

2) Strategy for Labelling Selected Data: Data collection consists of two phases:

Initial Phase: In order to get an initial estimate of the distribution of samples in the sample space, we start the procedure by observation. For the first N samples (100 samples in our setup; $\sim$25 hours of wearing the watch), we do not trigger any EMAs. At the end of this phase, we get an estimate of the distribution of samples in the samples space.

Query Phase: For samples after the initial phase ($N+1$ and above), we trigger the EMA (ask for labels) for a portion of samples. The probability of selecting each sample to be labeled is proportional to the number of previous samples (unlabeled) in its neighborhood. 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 ask the user for the label. For each region after we collect sufficient number of labeled samples, we stop collecting labels. However, the minimum probability of triggering the EMA for a sample is $P = 0.1$. This means if a sample 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 regions with higher densities.

We capture the stress levels of fourteen volunteers through self-reported questionnaires over periods of between 1-3 months. The total number of samples, along with the number of labeled samples for each user is presented in Table I.

[1]**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.

III. STRESS DETECTION

In this section, we explore the possibility of predicting stressful vs non-stressful moments based on the collected signals. We train our stress detection models on both personal and general datasets to evaluate the performance. For stress detection, we use several machine learning classification algorithms such as Multi Layer Perceptron (MLP), Random Forest (RF), k-Nearest Neighbors (kNN), Support Vector Machine (SVM), and XGBoost. MLP is a class of feed-forward neural networks that can be trained to do nonlinear classification and regression tasks. RF is an ensemble learning method for classification that operates by constructing a number of decision trees at training time and outputting the class that is the mode of the classes of the individual trees. kNN uses the k nearest points and takes a majority vote to predict the class of the sample. SVM finds hyper planes and partitions the sample space into different classes. XGBoost is an implementation of Gradient Boosted Decision Trees that is fast and performs well in classification tasks. We train each of these classifiers on our dataset and analyse their performance using machine learning methods and F1 score as the evaluation metric.

IV. RESULTS AND ANALYSIS

Detecting stress by only using PPG signals in everyday settings is a challenging task [5]. To evaluate the validity of our models, we perform K-Fold cross validation ($K = 5$). In K-Fold Cross Validation, we split the data into K equally sized segments (folds) and in each iteration use 1 fold for evaluation, and the other K-1 folds for training. The data is stratified prior to be split in K folds, to ensure each fold is a proper representative of the whole. The ML methods we used are introduced and explained in Section III. To evaluate the performance of each method on our collected dataset, we used Macro-F1 score[2]. The mean and standard deviation ($\mu \pm \sigma$) of the F1 scores over the K folds are presented in Table II. Based on these experiments, RF outperforms other methods for most cases (except for the first row).

A. Personalization

The bias in the physiological data (both the signals and the labels) can be different for personal or general datasets [5]. Therefore, we show the effect of personalization and how it improves the prediction accuracy on our collected dataset. To this end, we consider 3 participants from which more data is available (subjects S06, S10, S12). In the first step, we exclude the data from one subject (e.g. S06), train on the data from all other subjects, then test on half of the data from S06 (picked randomly). In the next step, in order to personalize the model, we use the other half of data from S06 and use it for training (along with the data from other users), and then test it on the first half of the data from S06. The results are reported in Table III. As can been seen from these results, personalization improves the prediction performance (macro-F1 score).

[2]F1 score is defined on each class separately. macro-F1 score is the average of F1 scores on all the classes (two here).

TABLE II
MEAN VALUE OF F1 SCORE FOR 5 FOLD CROSS VALIDATION (± STANDARD DEVIATION) FOR STRESS DETECTION BASED ON PPG FEATURES, BASELINE IS ALWAYS "NOT AT ALL" CLASS

Binary Classes	Number of Samples	5 fold Cross Validation, F1 Score				
		MLP	SVM	kNN	RF	XGBoost
a little bit (1) VS. baseline (0):	(605, 143)	**0.73 ± 0.06**	0.66 ± 0.03	0.72 ± 0.03	0.67 ± 0.04	0.72 ± 0.04
some (1) VS. baseline (0):	(299, 143)	0.70 ± 0.03	0.69 ± 0.04	0.66 ± 0.06	**0.71 ± 0.05**	0.70 ± 0.04
a lot or *extremely* (1) VS. baseline (0):	(72, 143)	0.68 ± 0.06	0.69 ± 0.13	0.69 ± 0.04	**0.76 ± 0.05**	0.73 ± 0.09
some, a lot or *extremely* (1) VS. *a little bit* or *not at all* (0):	(748, 371)	0.62 ± 0.04	0.60 ± 0.04	0.59 ± 0.03	**0.63 ± 0.02**	0.63 ± 0.04

TABLE III
EFFECT OF PERSONALIZATION ON STRESS PREDICTION PERFORMANCE. BEFORE AND AFTER ROWS ARE F1 SCORES BEFORE AND AFTER PERSONALIZATION FOR EACH USER.

		F1 Score on 50% of data from one user				
User	Personalization	MLP	SVM	KNN	RF	XGB
S06:	before	0.43	0.37	0.44	0.40	0.38
	after	**0.53**	**0.54**	**0.50**	**0.54**	**0.52**
S10:	before	0.58	0.63	0.60	0.612	0.60
	after	**0.62**	0.62	**0.63**	**0.616**	**0.61**
S12:	before	0.58	0.59	0.61	0.59	0.55
	after	**0.63**	**0.62**	**0.64**	**0.63**	**0.61**

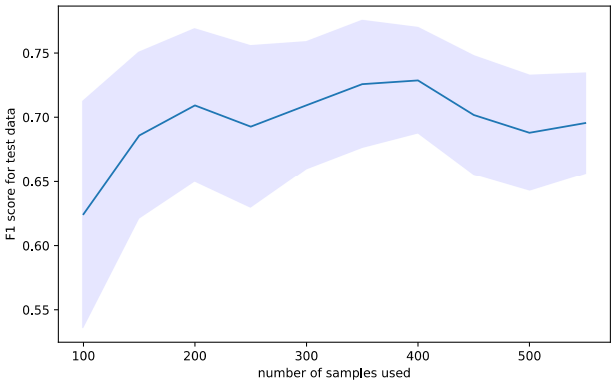

Fig. 2. Stress Prediction over size of training data

B. Improvement Over Size of Training Data

As we collect more labels from a user, the stress prediction can be performed more accurately. In order to show this process, we only use data from one user having for him a large number of labels is available (Subject S12). We also use random forest classifier for this experiment. We randomly separate 100 samples and use those for testing. We use the rest of the data in an incremental manner; first we train the model on 100 samples only, and then increase the training size. In each step, we test the trained model on the test data (all from S12). The improvement of prediction performance is presented in Figure 2. The model improves as we increase the subject's data size up to about 300 samples. For each step, we repeat the process (selection of test data and training data) 100 times, and the values in the figure show the mean and the standard deviation of the F1 score over all these 100 experiments.

V. CONCLUSIONS AND FUTURE WORK

Collecting photoplethysmogram (PPG) signals with enough labels collected from users in everyday settings is a challenging task. Our study used a Samsung Gear Sport smartwatch as a wearable device for data collection and utilized a method to improve the labeling procedure. The data were collected from fourteen active volunteers. We explored the possibility of detecting stressful vs non stressful moments (samples) using leave-samples-out validation based on PPG signals, in everyday settings. We analyzed the improvement of the trained classifier, as we personalize the classifier with samples from a certain subject. The results are promising: we achieved macro-F1 scores up to 76% for binary classification of stressful vs non stressful samples. This motivates our future work to utilize more advanced methods, possibly variants of active learning, in the labeling procedure. More informative labels will allow us to design classifiers that can possibly detect mental health conditions of users based on HRV from their biosignals and the type of user activity – promising to provide valuable tools for mental health professionals to better diagnose and treat stress and anxiety in a personalized way.

REFERENCES

[1] K. Glanz and M. D. Schwartz, "Stress, coping, and health behavior.," 2008.

[2] J. Bakker, M. Pechenizkiy, and N. Sidorova, "What's your current stress level? detection of stress patterns from gsr sensor data," in *2011 IEEE 11th international conference on data mining workshops*, pp. 573–580, IEEE, 2011.

[3] F. Firouzi *et al.*, "Internet-of-things and big data for smarter healthcare: From device to architecture, applications and analytics," *Future Generation Computer Systems*, vol. 78, pp. 583 – 586, 2018.

[4] R. Mieronkoski *et al.*, "The internet of things for basic nursing care—a scoping review," *International Journal of Nursing Studies*, vol. 69, pp. 78 – 90, 2017.

[5] H. J. Han *et al.*, "Objective stress monitoring based on wearable sensors in everyday settings," *Journal of Medical Engineering & Technology*, vol. 44, no. 4, pp. 177–189, 2020.

[6] E. K. Naeini *et al.*, "A real-time ppg quality assessment approach for healthcare internet-of-things," *Procedia Computer Science*, vol. 151, pp. 551 – 558, 2019.