Forecasting stress, mood, and health from daytime physiology in office workers and students

Forecasting stress, mood, and health from daytime physiology in office workers and students

Terumi Umematsu[1], Akane Sano[2 3], Sara Taylor[3], Masanori Tsujikawa[1], Rosalind W. Picard[3]

***Abstract*— We examine the problem of forecasting tomorrow morning's three self-reported levels (on scales from 0 to 100) of stressed-calm, sad-happy, and sick-healthy based on physiological data (skin conductance, skin temperature, and acceleration) from a sensor worn on the wrist from 10am-5pm today. We train automated forecasting regression algorithms using Random Forests and compare their performance over two sets of data: "workers" consisting of 490 days of weekday data from 39 employees at a high-tech company in Japan and "students" consisting of 3,841 days of weekday data from 201 New England USA college students. Mean absolute errors on held-out test data achieved 10.8, 13.5, and 14.4 for the estimated levels of mood, stress, and health respectively of office workers, and 17.8, 20.3, and 20.4 for the mood, stress, and health respectively of students. Overall the two groups reported comparable stress and mood scores, while employees reported slightly poorer health, and engaged in significantly lower levels of physical activity as measured by accelerometers. We further examine differences in population features and how systems trained on each population performed when tested on the other.**

I. Introduction

An accurate forecast of tomorrow's well-being might inspire people to make changes to their schedule today or tonight in order to improve their well-being tomorrow. Early detection indicators that one's well-being is getting worse may also enable new kinds of interventions to potentially prevent a series of bad stress or bad mood days from taking a turn into clinical depression or anxiety. Stress is well-known to increase susceptibility to infection and illness [1]. Self-reported health strongly relates to actual health and all-cause mortality [2]. Self-reported mood is strongly correlated to measures of depression [3]. The ability to forecast well-being levels, and identify what specifically changes them, could enable better self-management of behavioral choices, potentially preventing poor well-being and its damage to physical and mental health. The ability to model and forecast well-being could thus be immensely beneficial to society, especially if made in a privacy-sensitive, convenient and unobtrusive way.

This work is supported by the National Institute of Health (R01GM105018), the National Science Foundation (#1840167), the MIT Media Lab Consortium, Samsung Electronics and NEC Corporation. We thank our collaborators and study participants (https://snapshot.media.mit.edu/team/). The authors would like to thank Yoshiki Nakashima in NEC Corporation for collecting the office workers data with the experiments.

[1]NEC Corporation, Biometrics Research Laboratories. 1753 Shimonumabe, Nakahara-ku, Kawasaki, Kanagawa, Japan. {terumi,tujikawa}@nec.com

[2]Rice University, Department of Electrical and Computer Engineering akane.sano@rice.edu

[3]Massachusetts Institute of Technology, Media Lab.{sataylor, picard}@media.mit.edu

Previous work has shown that students' well-being (high/low stress, good/poor mood and good/poor health levels) tomorrow can be predicted with 78-82% classification accuracy based on today's physiological and behavioral data by using personalized machine learning models [4]. Previous work also showed that using 7 days of time series data with recurrent neural network (RNN) models can give acceptable results in well-being prediction without building personalized prediction models for forecasting students' tomorrow's high/low stress [5]. Other work has shown that using only daytime (10am-5pm) physiology data in 7 days of time series with long short-term memory neural network models (LSTM) can forecast the next-day's students' stress, mood, and health [6]. However, these works focus on student populations instead of workforce populations, despite the growing interest in this space [7]–[10]. It has not yet been examined whether daily stress, mood, and health levels for office workers can be accurately forecast using only passively collected physiology data over the workday, or how the results compare to those in a very different population (e.g. students).

In this work, we investigate the hard problem of forecasting the level of tomorrow's self-reported wellbeing (stress, mood, and health, reported each morning on a scale from 0 to 100) using only daytime, passively-collected physiology data from today. Daytime is defined here as 10am-5pm, a period of time when both office workers and the group of college students were generally awake and active. We further restrict the algorithms to use only physiological features (skin conductance, skin temperature, and acceleration). The problem is especially challenging because, in order to preserve nighttime privacy, we do not use any data from tonight when predicting tomorrow morning's wellbeing.

This paper makes new contributions expanding automated means of forecasting well-being for office workers. Our results also provide new insights into how the physiological features and the performances of the automated methods compare across the student and office-worker populations.

II. Data

A. Data Sets

(1) Workers' data

A total of 39 workers from one Japanese IT company (number of employees: >20,000) in 2017 collected physiological and behavioral survey data over a 30-day period. The participants ranged in age from 20s-50s and were from the following departments: R&D 50%, developer 28%, sales

10%, planing 7%, and system engineer 5%. Each participant wore a wearable sensor during their working time during the weekdays. Stress, mood, and health scores were collected each morning at the start of the participant's working day (around 9am). A total of 490 days of complete daytime data was collected.

(2) Students' data

The students' data in this experiment came from the study Sleep, Networks, Affect, Performance, Stress, and Health using Objective Techniques (SNAPSHOT) [11], which gathered 30-day multi-modal data, including physiological, mobile phone, and behavioral survey data from college students in one US university during 2015-2017. As in the workers' dataset, stress, mood, and health scores were collected every morning. The study participants obtained compensation based on their contribution to the study. In this work, we removed days of data that were missing a self-reported score and we removed weekend data in order to have a more similar dataset to the workers data. We used a total of 3,841 days of daytime data from weekdays from 201 students.

B. Self-reported Survey for Ground-truth Scores and Checking Data Distribution

Self-reported stress, mood, and health scores were collected every morning, using self-reported scores from 0 (stressed out) - 100 (calm), 0 (sad) - 100 (happy), and 0 (sick) - 100 (healthy), respectively. These scores were used as the ground-truth labels for the forecasts. For checking data distribution, participants filled out a few minutes of survey about their daily behaviors every morning at the same time. They self-reported the duration of some activities, including sleep, active time (academic and study activities for students, working duration for workers), and exercise. We also collected the Perceived Stress Scale questionnaire [12] (PSS-10 score: 0 (low stress) - 40 (high stress)) for comparing distributions.

C. Physiology Feature Calculation

We computed 42 daily physiology features for both workers' and students' data in the same way. The physiological measurements were collected by wrist-worn Empatica E4 sensors from office workers and by wrist-worn Affectiva Q sensors from students; each sensor records electrodermal activity (EDA) measured as skin conductance (SC), skin temperature (ST), and 3-axis acceleration (ACC). The sampling rate of E4 sensors is 4 Hz for EDA and SC, and 32 Hz for ACC, and the sampling rate of Q sensors is 8 Hz for EDA, SC, and ACC. We defined data during 10am-5pm (10-17H) as "daytime" data as all workers and most students were active during this timeframe, and we used only "daytime" data for experiments. EDA, acceleration and ST were collected to measure sympathetic nervous activity, physical activity, circadian rhythm, and stress responses [13]–[15]. Following [16] and [4], for each time period the following sets of features were computed: EDA Peak features (for all detected peak features and for only non-artifact peaks [17]), SC level features, accelerometer features, temperature features, and various combinations of the three physiological data streams. All physiology features are explained in Table I.

III. Daily Well-being Forecasting Experiments

A. Experimental Conditions

We examine how accurately the previous days' physiology data using only a daytime day's data can forecast a next-day's morning well-being level. Specifically, we learn $p(y_{t+1}|x_t)$, the probability of the person's well-being given the previous daytime days' data, where x_t is the physiology data collected from wearable sensors on day t, and y_{t+1} is the next-day self-reported well-being scores.

B. Regression Labels

We framed the problem as a regression: for forecasting tomorrow morning's well-being (stress, mood, and health) using today's physiology features (II.C). The three daily labels are the values from 0 to 100 given in the morning for each of the stress, mood, and health scales.

C. Regression Methods

We used Random Forests for regression. Random Forests are an ensemble learning method using a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest [18]. Parameters such as the number of trees, the maximum depth of the tree, and the minimum number of samples required to split an internal node were selected by grid search. The full dataset was used in a five-fold cross validation with 80% of the data for training and validating the models, and 20% for testing each fold. Specifically, within the training and validation set, we used 80% of the dataset for training and 20% as validation and selected the hyperparameters (the number of trees in the forest and the maximum depth of the tree) that yielded the highest accuracy on the validation set. The days in the test set were kept completely independent of the training and validation data. The whole algorithm was implemented using scikit-learn library and Python 3.5.4.

D. Evaluation metrics

We used Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and correlation as evaluation metrics. For evaluating accuracy we computed the average and the Standard Deviation (SD) of the test set for the five folds. Specifically, let n be the number of samples and e_t be the error between forecasting results and labels, respectively.

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|e_t|, RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}e_t^2} \tag{1}$$

Using the set-up above, we compare the accuracies of using the two training data.

TABLE I

PHYSIOLOGY FEATURES [4], [16], [17]

Category	Feature's name	Explanation of Features
		EDA Peaks Features
EDA	sumAUC	the sum of the AUC of all peaks where amplitude of peak is calculated as difference from base tonic signal
	sumAUCFull	sum of AUC of peaks where amplitude is calculated as difference from 0
	medianRiseTime	median rise time of peaks (seconds)
	medianAmplitude	median amplitude of peaks (μS)
	countPeaks	number of peaks detected
	sdPeaks30min	compute number of peaks per 30 minute epoch, take SD of this signal
	medPeaks30min	compute number of peaks per 30 minute epoch, take median of this signal
	percentHighPeak	percentage of signal containing 1 minute epochs with greater than 5 peaks
		EDA Peaks without Artifact Features
EDA without artifact (EDA noA)	sumAUCNoArtifact	sumAUC without artifact
	sumAUCFullNoArtifact	sumAUCFull without artifact
	medianRiseTimeNoArtifact	medianRiseTime without artifact
	medianAmplitudeNoArtifact	medianAmplitude without artifact
	countPeaksNoArtifact	countPeaks without artifact
	sdPeaks30minNoArtifact	sdPeaks30min without artifact
	medPeaks30minNoArtifact	medPeaks30min without artifact
	percentMedPeakNoArtifact	percentMedPeak without artifact
	percentHighPeakNoArtifact	percentHighPeak without artifact
		Skin conductance level (SCL) Features
SCL	sclPercentOff	percentage of period where sensor was off
	sclMaxUnnorm	max level of un normalized EDA signal
	sclMedUnnorm	median of normalized EDA signal
	sclMeanUnnorm	mean of un-normalized EDA signal
	sclMedianNorm	median of z-score normalized EDA signal
	sclSDnorm	standard deviation of z-score normalized EDA signal
	sclMeanDeriv	mean derivative of z-score normalized EDA signal ($\mu S/second$)
		Accelerometer Features
ACC	stepCount	number of steps detected
	meanMovementStepTime	average number of samples (at 8Hz) between two steps
	stillnessPercent	percentage of time the person spent nearly motionless
		Accelerometer Weighted Peak Features
ACC weighted EDA	sumStillnessWeightedAUC	weight the peak AUC signal by how still the user was every 5 minutes and sum
	sumStepsWeightedAUC	weight the peak AUC signal by the step count over every 5 minutes and sum
	sumStillnessWeightedPeaks	multiply the number of peaks every 5 minutes by the amount of stillness during that period
	maxStillnessWeightedPeaks	the max value for the number peaks * stillness for any five minute period
	sumStepsWeightedPeaks	divide number of peaks every five minutes by step count and sum
	medStepsWeightedPeaks	average value for the number of peaks / step count every 5 mins
		Temperature Weighted EDA peaks Features
ST weighted EDA	sumTempWeightedAUC	sum of peak AUC divided by the average temp every 5 mins
	sumTempWeightedPeaks	number of peaks divided by the average temp every 5 mins
	maxTempWeightedPeaks	the maximum number of peaks in any 5 minute period divided by the average temp
		Skin Temperature Features
ST	maxRawTemp	the maximum of the raw temperature signal (°C)
	minRawTemp	the minimum of the raw temperature signal (°C)
	sdRawTemp	the standard deviation of the raw temperature signal (°C)
	medRawTemp	the median of the raw temperature signal (°C)
		Accelerometer Weighted Skin Temperature
ACC weighted ST	sdStillnessTemp	the standard deviation of the temperature recorded during periods when the person was still
	medStillnessTemp	the median of the temperature when the person was still

IV. RESULTS AND DISCUSSION

A. Group-Level Data Distribution

First we examine the results of the surveys described in section II.B: Table II shows the mean and standard deviation (SD) of workers' and students' daily well-being and PSS-10 scores (pre: before 30-days data collection, post: after 30-days data collection). Stress and mood scores are not significantly different for the groups (Welch's t test), but health and pre-PSS scores are different, with workers less healthy and initially more stressed. Table III shows the mean and standard deviation for workers' and students' self-reported durations of time in bed, active, and exercising. All of these durations differed significantly between the groups (p <0.05, Welch's t test). On average, workers spent more time sleeping and working, and less time exercising than did students.

TABLE II

MEAN(SD) OF DAILY WELL-BEING AND PSS SCORES

	Workers	Students	p-value
Stress score	53.96 (21.26)	53.99 (25.63)	0.98
Mood score	56.90 (17.14)	61.03 (22.47)	8.83
Health score	60.09 (21.93)	64.60 (25.58)	<0.05
PSS-10 score (pre)	17.77 (6.21)	14.73 (7.09)	<0.05
PSS-10 score (post)	16.62 (5.67)	16.13 (7.44)	0.72

B. Well-being Forecasting Results

We examine how accurately (using MAE, RMSE, and correlation) the models using workers' or students' data forecast next-day stress, mood, and health in section III. In Tables IV, V, and VI, we first confirm (as expected) that higher accuracy is obtained testing on students after training on students, and similarly for workers. The MAE

TABLE III

MEAN (SD) OF SLEEP, ACTIVE, AND EXERCISES DURATION (IN MINUTES)

	Workers	**Students**	**p-value**
Bed time duration[mins]	360.37 (90.32)	337.41 (174.50)	<0.05
Active time duration[mins]	533.19 (119.18)	Academic: 154.84 (135.47)	<0.05
		Study: 166.20 (165.70)	<0.05
Exercise time duration[mins]	16.35 (42.18)	29.20 (64.78)	<0.05

TABLE IV

STRESS FORECASTING ACCURACY (MAE (RMSE), CORR)

		Train	
		Workers	Students
Test	Workers	**13.47 (19.95), 0.37**	17.56 (22.50), -0,04
	Students	22.72 (27.45), -0.07	**20.28 (24.76), 0.18**

TABLE V

MOOD FORECASTING ACCURACY (MAE (RMSE), CORR)

		Train	
		Workers	Students
Test	Workers	**10.80 (14.09), 0.67**	15.64 (20.07), -0.22
	Students	25.66 (30.73), -0.10	**17.81 (22.20), 0.13**

TABLE VI

HEALTH FORECASTING ACCURACY (MAE (RMSE), CORR)

		Train	
		Workers	Students
Test	Workers	**14.41 (18.51), 0.45**	18.32 (21.67), -0.05
	Students	30.66 (35.84), -0.04	**20.43 (24.47), 0.22**

using Random Forest are 13.47 for stress, 10.80 for mood, and 14.41 for health, using physiology features from workers. These results on the office workers are consistently better than the model trained on students, which obtained in the best case an MAE of 20.28 for stress, 17.81 for mood, and 20.43 for health when it was tested on the students. One possible reason for higher accuracy with the workers is that their data may be more homogeneous than the student data, given lower variances and less change in their PSS-10 score as shown in Table II, and lower physical activity levels during the daytime. In addition, all the office workers were from the same culture (Japanese) while the students at this university come from many diverse cultures. Previous studies showed higher accuracy for predicting students' stress, mood, and health using the same students' data and deep-learning methods [19]–[21] while we used Random Forest, one of interpretable machine learning methods, that allowed us to interpret feature importance.

C. Feature Importance for Worker and Student Models

We computed the top 10 features for each Random Forest model of section IV.B, which we list in Figs 1 and 2. Features with higher weights indicate a stronger influence on forecasting stress, mood, and health. High feature importance of both groups are ACC features such as step count related to increased physical activity. All six models included two accelerometer based features: stepcount and stillnessPercent.

In Table VII we show the mean and standard deviation (SD) of the top three important features in each model. All three features show a significant difference between the two groups. Over weekdays 10am-5pm, office workers showed lower levels of step counts while students showed higher levels (workers: average 1,699, SD 1,145, and students: 3,428, 1,956, p <0.05, Welch's t test). In addition, we found that for weekdays 10am-5pm, the office workers showed lower levels of maximum skin temperature compared to the students (office workers: average 31.3°C, SD 2.0, and students: 35.4°C, 2.5, p <0.05, Welch's t test), and the office workers showed a higher average number of samples between two steps (ACC features: meanMovementSteptime, office workers: average 356.25, SD 248.63, and students: 230.33, 256.62, p <0.05, Welch's t test). As is generally true for different populations of data, using different group models for training and testing gives lower accuracy than using the same group model, as shown in Tables IV, V, and VI.

TABLE VII

MEAN AND STANDARD DEVIATION (SD) OF THE TOP THREE OF IMPORTANT FEATURES IN EACH MODELS

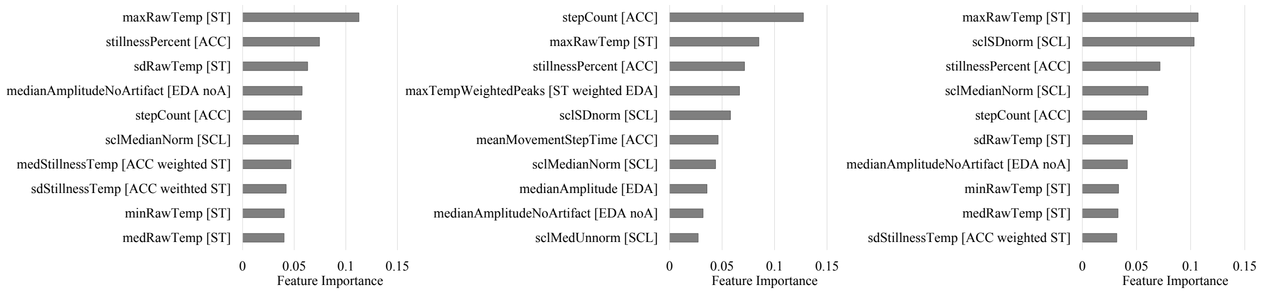
	Workers		**Students**		
	Mean	SD	Mean	SD	p-value
maxRawTemp [ST]	31.33	2.00	35.42	2.52	<0.05
stepCount [ACC]	1699.32	1144.72	3428.44	1956.48	<0.05
meanMovementStepTime [ACC]	356.25	248.63	230.33	256.62	<0.05
stillnessPercent [ACC]	0.70	0.15	0.56	0.18	<0.05
sclSDnorm [SCL]	0.69	0.57	0.53	0.43	<0.05
sclMedianNorm [SCL]	0.57	0.38	-0.05	0.38	<0.05
medianAmplitudeNoArtifact [EDA noA]	0.30	0.10	0.23	0.14	<0.05
sclMeanUnnorm [SCL]	2.53	0.55	0.41	0.57	<0.05
sdRawTemp [ST]	0.88	3.30	1.82	0.78	<0.05
sclMedUnnorm [SCL]	2.16	3.16	0.29	0.50	<0.05

V. CONCLUSIONS AND FUTURE WORK

In this paper, we developed an automated forecast for workers' daily well-being scores using physiological data that has previously been shown to accurately estimate well-being in students. We also focused on making a forecast for tomorrow morning's values based on only the daytime 10am-5pm physiological data today. The experimental results show that the daily well-being, measured as mood, stress, and overall physical health for workers can be forecast on a scale from 0 to 100 with MAE of less than 15 points. The main difference in physiology features (measured on the wrist) between workers and students are related to acceleration (e.g., office workers having lower step count) which is likely related to differing environments and behavioral patterns.

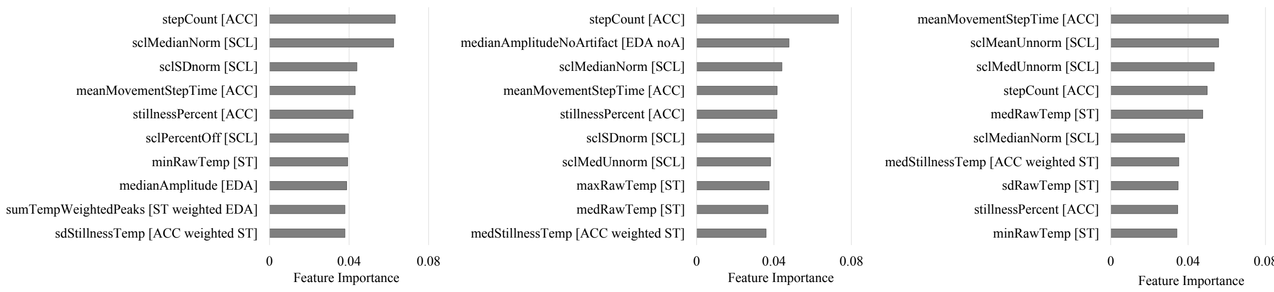
While this work has expanded automated forecasting abilities, this work has several limitations. The two datasets are collected on office workers in a Japanese IT company and New England college students and might not generalize to other populations. Office workers at an IT company were chosen since, like the college students they are merit-driven and work with technology a large part of the day. More populations need to be studied before we can draw general conclusions about how much these models need to be customized to different cultures, ages, and activity levels. However, this work appears to be the first to expand daily well-being forecasting models using physiology data to office workers.

In future work, we plan to collect more data for not only office workers but also other workers such as field workers, and with longer monitoring per person, it might be possible to build more accurate forecasting models. Further, we plan to examine transfer learning and other deep learning methods to improve well-being forecasting accuracies by



(a) Stress (b) Mood (c) Health

Fig. 1. Features Importance of workers' model



(a) Stress (b) Mood (c) Health

Fig. 2. Features Importance of students' model

using previously collected data in novel populations. In addition, we will consider adaptive methods to fill in missing data with time series information [22]. Finally, this work has focused on the forecasting and has not addressed several aspects of the problem of what to do with the forecasts – how to help people identify the best behaviors to change and how to support them in making those changes. These are important challenges to solve before closing the feedback loop with the participants. Nevertheless, this work shows that earlier work, limited to forecasting in students, can indeed be expanded to provide well-being forecasting in office workers, and in fact can even work more accurately in this important population.

REFERENCES

[1] S. Cohen *et al.*, "Psychological Stress and Susceptibility to the Common Cold," *New England Journal of Medicine*, vol. 325, no. 9, pp. 606–612, 1991.

[2] K. Abiola *et al.*, "Does the perception that stress affects health matter? the association with health and mortality," *Health Psychology*, vol. 31, no. 5, pp. 677–684, 2012.

[3] H. Cheng *et al.*, "Personality, self-esteem, and demographic predictions of happiness and depression," *Personality and individual differences*, vol. 34, no. 6, pp. 921–942, 2003.

[4] S. A. Taylor *et al.*, "Personalized Multitask Learning for Predicting Tomorrow's Mood, Stress, and Health," *IEEE Transactions on Affective Computing*, no. 99, pp. 1–14, 2017.

[5] T. Umematsu *et al.*, "Improving students' daily life stress forecasting using lstm neural networks," *IEEE EMBS International Conference on Biomedical & Health Informatics*, pp. 1–4, 2019.

[6] T. Umematsu *et al.*, "Daytime data and lstm can forecast tomorrow's stress, health, and happiness," *EMBC.*, pp. 2186–2190, 2019.

[7] T. G. Vrijkotte, L. J. van Doornen, and E. J. de Geus, "Effects of Work Stress on Ambulatory Blood Pressure, Heart Rate, and Heart Rate Variability," *Hypertension*, vol. 35, pp. 880–886, 2000.

[8] D. Carneiro, P. Novais, J. C. Augusto, and N. Payne, "New Methods for Stress Assessment and Monitoring at the Workplace," vol. 14, no. 8, pp. 1–18, 2015.

[9] S. Koldijk, M. A. Neerincx, and W. Kraaij, "Detecting Work Stress in Offices by Combining Unobtrusive Sensors," *IEEE Transactions on Affective Computing*, vol. 9, no. 2, pp. 227–239, 2018.

[10] Y. Nakashima *et al.*, "Improvement in chronic stress level recognition by using both long-term and short-term measurements of physiological features," *EMBC*, 2018.

[11] A. Sano *et al.*, "Identifying objective physiological markers and modifiable behaviors for self-reported stress and mental health status using wearable sensors and mobile phones: Observational study," *J Med Internet Res*, vol. 20, no. 6, p. e210, 2018.

[12] S. Cohen *et al.*, "A global measure of perceived stress," *Journal of health and social behavior*, pp. 385–396, 1983.

[13] W. Boucsein, *Electrodermal activity*. Springer Science & Business Media, 2012.

[14] C. G. Scully *et al.*, "Skin surface temperature rhythms as potential circadian biomarkers for personalized chronotherapeutics in cancer patients," *Interface Focus*, vol. 1, no. 1, pp. 48–60, 2011.

[15] K. A. Herborn *et al.*, "Skin temperature reveals the intensity of acute stress," *Physiology & Behavior*, vol. 152, pp. 225–230, 2015.

[16] S. A. Taylor, "Characterizing Electrodermal Responses during Sleep in a 30-day Ambulatory Study," *MIT, Master's Thesis*, 2016.

[17] S. A. Taylor *et al.*, "Automatic identification of artifacts in electrodermal activity data," *EMBC*, pp. 1934–1937, 2015.

[18] L. Breiman, "Random forests," *Machine Learning 45*, pp. 5–32, 2001.

[19] N. Jaques *et al.*, "Predicting tomorrow's mood, health, and stress level using personalized multitask learning and domain adaptation," *Journal of Machine Learning Research*, vol. 66, pp. 17–33, 2017.

[20] B. Li *et al.*, "Toward end-to-end prediction of future wellbeing using deep sensor representation learning," *ACII*, pp. 253–257, 2019.

[21] H. Yu *et al.*, "Personalized wellbeing prediction using behavioral,physiological and weather data," *IEEE International Conference on Biomedical and Health Informatics*, 2019.

[22] N. Jaques *et al.*, "Multimodal Autoencoder : A Deep Learning Approach to Filling In Missing Sensor Data and Enabling Better Mood Prediction," *ACII*, pp. 1–7, 2017.