Sensors and Materials, Vol. 32, No. 2 (2020) 703–722
MYU Tokyo

S & M 2131

Effect of Person-specific Biometrics in Improving Generic Stress Predictive Models

Kizito Nkurikiyeyezu,* Anna Yokokubo, and Guillaume Lopez

Aoyama Gakuin University, Graduate School of Science and Engineering,
5-10-1 Fuchinobe, Chuo-ku, Sagamihara-shi, Kanagawa-ken 252-5258, Japan

(Received October 4, 2019; accepted November 26, 2019)

Keywords: continuous stress monitoring, physiological computing, heart rate variability, electrodermal activity, smart buildings

Because stress is subjective and is expressed differently from one person to another, generic stress prediction models (i.e., models that predict the stress of any person) perform crudely. Only person-specific models (i.e., ones that predict the stress of a preordained person) yield reliable predictions, but they are not adaptable and are costly to deploy in real-world environments. For illustration, in an office environment, a stress monitoring system that uses person-specific models would require the collection of new data and the training of a new model for every employee. Moreover, once deployed, the models would deteriorate and need expensive periodic upgrades because stress is dynamic and depends on unforeseeable factors. We propose a simple, yet practical and cost-effective calibration technique that derives an accurate and personalized stress prediction model from physiological samples collected from a large population. We validate our approach on two stress datasets. The results show that our technique performs much better than a generic model. For instance, a generic model achieved only 42.5 ± 19.9% accuracy. However, with only 100 calibration samples, we raised its accuracy to 95.2 ± 0.5%. We also propose a blueprint for a stress monitoring system based on our strategy, and we debate its merits and limitations. Finally, we made our source code and the relevant datasets public to allow other researchers to replicate our findings.

1. Introduction

Occupational stress is well researched,[1–3] not only due to its pernicious effect on people's health but also due to the economic benefits of keeping in check the stress level of employees. Although a small amount of stress is benign and even beneficial because it provides the necessary impetus to survive the tribulations of the modern workplace,[4,5] chronic stress (i.e., enduring stress) has detrimental repercussions. Physiological and psychological disorders,[6] job-related tensions,[10] and general deterioration of health are just a few examples of its adverse outcomes. Furthermore, stress is liable for significant economic losses because stressed workers have suboptimal productivity, are prone to higher job absenteeism and presenteeism, and are disproportionately predisposed to sickness.[7,8]

*Corresponding author: e-mail: kizito@wil-aoyama.jp
https://doi.org/10.18494/SAM.2020.2650

ISSN 0914-4935
https://myukk.org/

Consequently, the importance of overcoming stress at work is essential to the well-being of the workers and the bottom line of any business. Nevertheless, at the moment, there is no mainstream real-world stress monitoring system.[9] The most reliable stress monitoring strategies rely on directly measuring the level of stress-inducing hormones [e.g., salivary and cortisol concentration in sweat[10,11]] and on psychological evaluations performed by psychologists. However, these procedures are neither suitable nor feasible for continuously monitoring stress in workplaces because they are obtrusive and are carried out sporadically. Moreover, in the case of physiological evaluations, people are reluctant to reveal their work stress honestly.[12] However, stress spawns detectable physiological, psychological, and behavioral changes that can be used for automatic stress recognition.[1,3] For example, acute stress decreases a person's heart rate variability (HRV) and parasympathetic activation.[13] Besides, there is plentiful research evidence that it is plausible to indirectly monitor stress using physiological signals such as the electrodermal activity (EDA),[14] HRV,[15,16] electroencephalogram (EEG),[17] and electromyogram (EMG).[18]

Although there are many publications[1–3,19] on automatic stress prediction, at the moment, aside from a few niches and non-scientifically proven consumer products, there is no effective system that automatically and unobtrusively monitors people's stress in real-world environments.[9] On the one hand, some of the proposed approaches (e.g., EEG-based stress monitoring) are impractical because they are too obtrusive. On the other hand, the most precise approaches (e.g., Refs. 19–21) predict stress using a fusion of multiple sensor data (e.g., audio, video, computer logging, posture, facial expression, and physiological features). These methods, however, raise technical, privacy, and security challenges (e.g., user's computer keystroke logging, video recording, and speech recording), and are therefore inconvenient to deploy in real-world settings because of company-wide computer security policies or international workplace privacy regulations. Finally, the most practical and unobtrusive stress monitoring methods (e.g., Refs. 22 and 23)—which are mostly based on physiological signals that are recordable on a person's wrist [e.g., photoplethysmograms (PPGs) and EDA—are not yet widely available to the general public despite their potential economic and health benefits. The lack of viable stress monitoring products, despite the extensive research on occupational stress, the availability of enabling technologies (e.g., smartphones with on-wrist HRV and EDA sensors), and the immense economic and health benefits such products would bring, begs the question of why this is the case.

A recent review article on affect and stress recognition[2] scrutinized the published literature and noted the striking discrepancy between the accuracy of person-specific stress prediction machine learning (ML) models (i.e., ML models that predict the stress of a specific person) and person-independent ML models (i.e., generic ML models that predict the stress of any person). The article underscores that person-specific ML models (e.g., Refs. 15, 20, and 23–25) achieved excellent prediction accuracy. Nevertheless, their predictions are person-specific—that is, the ML models would not generalize well in predicting the stress of yet unseen people; therefore, they cannot be used in creating mass-market stress monitoring products. In contrast, pragmatic person-independent solutions (e.g., Refs. 22, 26, and 27) generally have a much lower stress prediction accuracy; accordingly, they are also a poor choice for creating mass-market stress

monitoring products. For example, 95.0% emotion recognition accuracy using person-specific ML models was achieved in Ref. 27; however, the same approach resulted in only 70% accuracy when applied to a person-independent classification model. In a similar manner, the authors in Ref. 24 conducted experiments to monitor stress in daily work and found that ML models that use people's physiology to predict stress are highly person-dependent. Person-specific ML models achieved 97% accuracy, but the generic ones only achieved 42% accuracy. Their results resemble those in Ref. 20, which achieved 90.0% accuracy when using person-specific stress classification models. However, when the same approach was applied to predict the stress of new subjects, its performance ebbed decreased to 58.8 ± 11.6% accuracy.

These modest outcomes are expected. For example, the authors in Ref. 28 argued that, when people's physiological differences are not accounted for, the ML stress prediction models performed no better than a model with no learning capability. First, stress is intrinsically idiosyncratic and depends on a person's uniqueness (e.g., his/her genetics) and coping ability.[29] Second, there is incontrovertible evidence that there are gender differences in how people respond to stress[30] and that men and women have different feelings about stress because women tend to express a higher level of stress on self-report questionnaires.[31] Third, a stressor that produces stress in one person will not necessarily trigger the same stress response in a different person.[32] Finally, for the same person, there is a significant day-to-day variability in cortisol awakening response, which may affect how that person responds to stress.[33] As a result, a practical stress monitoring scheme needs to take into account inter-individual and intra-individual differences, gender, the temporal variability of human stress, and many other factors that influence how humans react to stress. The state-of-the-art stress monitoring strategies (e.g., Ref. 34) use person-specific ML models. Unfortunately, this method is not realistic for creating a real-world product. A stress monitoring system that uses this approach would be costly (e.g., collecting and training ML stress prediction models for every user of the system) and would require expensive recurrent updates because stress is innately dynamic.

In recent studies, diverse methods have been proposed to improve the performance of generic stress prediction models. The most straightforward methods use normalization techniques (e.g., range normalization, standardization, baseline comparison, and Box–Cox transformation) to reduce the impact of inter-individual variability while preserving the differences between stress classes.[28,35] The normalization improves the performance of generic models but always underperforms compared with the person-specific ones. Furthermore, as noted in Chap. 5 of Ref. 36, the normalization process is multifaceted and depends on trial and error methods. An alternative strategy is to predict stress based on clusters of similar users.[20,36] These techniques are important contributions to producing an effective stress monitoring system. However, they also perform inadequately compared with person-specific models. Moreover, these methods would likely prove too complex to use in real-world settings because they are sensitive to the number of clusters[36] and, given that many factors influence a person's stress,[37] it is not clear what are the criteria for similarity to create clusters of similar users.

In this paper, we propose a hybrid and cheaper-to-deploy stress prediction method that incorporates tiny person-specific physiological calibration samples into a much larger generic sample collected from a large group of people. The proposed method hinges on the premise

that all humans share a hormonal response to stress,[38] but that a person's unique factors such as gender,[30] genetics,[39] personality,[40] weight,[41] and coping ability[29] determine how the person reacts to stress. Hence, we hypothesize that it may be possible to reuse generic samples collected from many people as a starting point for creating a personalized and more effective model. To confirm these assumptions, we tested this strategy on two major stress datasets. Our results show a substantial improvement in the performance of stress prediction models even when we used only 100 calibration samples. In summary,

(i) for each subject in the datasets, we trained and validated *n* person-specific regression and classification stress prediction models using a 10-fold cross-validation (CV) approach. The result shows that, for all subjects, the classification models achieved greater than 95% classification accuracy and that the regression models had a near-zero mean absolute error (MAE).

(ii) We used leave-one-subject-out cross-validation (LOSO-CV) to assess the performance of generic stress prediction models. All models performed poorly (e.g., 42.5 ± 19.9% accuracy, 14.0 ± 7.9 MAE, on one dataset) compared with person-specific models and there was a wide performance variation between the subjects.

(iii) We devise a hybrid technique that derives a personalized person-specific-like stress prediction model from samples collected from a large population and discussed how it could be used to develop a real-world continuous stress monitoring system in, for example, intelligent buildings.

2. Methods

2.1 Stress datasets

We used two stress datasets to conduct this study. The first dataset—the SWELL dataset[42]—was collected at Radboud University. This dataset is a result of experiments conducted on 25 subjects doing office work (e.g., writing reports, making presentations, reading e-mails, and searching for information) who were exposed to quintessence typical work stressors (e.g., being unexpectedly interrupted by an urgent e-mail, and pressure to complete work in a limited time). During the experiment, the researchers recorded the subjects' computer usage patterns, facial expressions, body postures, and electrocardiogram (ECG), and EDA signals. The participants experienced three different working conditions:

1) No stress—The participants performed the assigned tasks for a maximum of 45 min.
2) Time pressure—The time given to these participants to finish their tasks was reduced to two-thirds of the required duration under the no-stress condition.
3) Interruption—The participants received interrupting e-mails in the middle of their assigned tasks. Some e-mails were relevant to their tasks, and the participants were requested to take specific actions. Other e-mails were irrelevant, and the participants did not need to take any action.

At the end of each experimental trial, each participant's perceived stress was assessed using a variety of self-report questionnaires, including the NASA Task Load Index (NASA-TLX).[43]

In this study, we focus on the NASA-TLX because it indicates a person's mental load based on a weighted average of multidimensional ratings (in terms of mental demand, physical demand, temporal demand, effort, performance, and frustration) and is the standard method in assessing subjective workload.

The second dataset—the WESAD dataset[26]—was collected by researchers from Robert Bosch GmbH and the University of Siegen in Germany. The dataset includes physiological (EDA, ECG, EMG, respiration signal, and skin temperature) and acceleration signals that the researchers collected from 15 subjects exposed to the following three affective stimuli:

(1) Baseline condition—The baseline condition aimed at generating a neutral affective state onto the participants and lasted for 20 min.
(2) Amusement condition—The subjects watched funny video clips. Each video clip was followed by a brief (5 s) neutral condition. The amusement condition lasted 392 s.
(3) Stress conditions—The participants were subjected to the Trier Social Stress Test (TSST)[44] and asked to give a 5 min public speech and to count down from 2023 in multiples of 17. If the subject made an error, he/she was requested to start over.

The amusement and stress conditions were each followed by a meditation period to "de-excite" the participants back to the baseline condition. Throughout the experiment, the participants provided five self-reports, including the Short Stress State Questionnaire (SSSQ),[45] which was used to determine the type of stress (i.e., worry, engagement, or distress) that was prevalent in the participants.

2.2 Feature extraction

We extracted HRV and EDA features from the two datasets. We computed the HRV features according to the standards and algorithms proposed by the Task Force of the European Society of Cardiology.[46] Each HRV feature (Table 1) was computed on a 5 min moving window as follows: first, we extracted an inter-beat interval (IBI) signal from the peaks of the ECG signal of each subject. Then, we computed each HRV index on a 5 min IBI array. Finally, a new IBI sample was appended to the IBI array, while the oldest IBI sample was removed from the beginning of the IBI array. The new resulting IBI array was used to compute the next HRV index. We repeated this process until the end of the entire IBI array. Likewise, for the EDA signals, the raw EDA signal was first filtered by a 4 Hz fourth-order Butterworth low-pass filter and then smoothed with a moving average filter. Next, we computed the EDA features (Table 1) on the 10 min moving window signal extracted from various EDA attributes of the skin conductance response (SCR).

All the resulting datasets—especially the WESAD datasets—are inherently unbalanced because their experimental protocols dictated different durations. We downsampled the datasets by randomly discarding some samples from the majority classes to make the datasets balanced, thus, preventing the majority classes from overshadowing the minority classes. Furthermore, for the WESAD dataset, we removed all samples corresponding to the amusement condition because its duration was almost as short as the sliding window we would have used to compare the feature.

Table 1
HRV and EDA features.

HRV	Time domain	Mean, median, standard deviation, skewness, and kurtosis of all RR intervals
	RMSSD	Root mean square of the successive differences
	SDSD	Standard deviation of all intervals of differences between adjacent RR intervals
	SDRR_RMSSD	Ratio of SDRR to RMSSD[56]
	pNNx	Percentage of number of adjacent RR intervals differing by more than 25 or 50 ms
	SD1, SD2	Short- and long-term Poincaré plot descriptor of HRV
	RELATIVE_RR[a]	Time domain features (e.g., mean, median, SDRR, and RMSSD) of the relative RR
	VLF, LF, HF	Very low (VLF), low (LF), high (HF) frequency band in the HRV power spectrum
	LF/HF	Ratio of low (LF) to high (HF) HRV frequencies
EDA	Time domain	Mean, max, min, range, kurtosis, and skewness of the SCR
	Derivatives	Mean and standard deviation of the first and derivatives of the SCR
	Peaks	Mean, max, min, and standard deviation of the peaks
	Onset	Mean, max, min, and standard deviation of the onsets[57]
	ALSC[b]	Arc length of the SCR
	INSC[c]	Integral of the SCR
	APSC[d]	Normalized average power of the SCR
	RMSC[e]	Normalized root mean square of the SCR

[a] $REL_{RR_i} = 2\left[\frac{RR_i - RR_{i-1}}{RR_i + RR_{i-1}}\right], i = 2,\ldots,N$

[b] $ALSC = \sum_{n=2}^{N}\sqrt{1 + \left(r[n] - r[n-1]\right)^2}$

[c] $INSC = \sum_{n=1}^{N}\left|r[n]\right|$

[d] $APSC = \frac{1}{N}\sum_{n=1}^{N} r[n]^2$

[e] $RMSC = \sqrt{\frac{1}{N}\sum_{n=1}^{N} r[n]^2}$

2.3 Feature engineering

An inspection of the histogram plots of the features computed in Sect. 2.2 revealed that the data distribution of most features is skewed. While this may not be an issue for some ML algorithms, in other cases, the distribution of the features is critical. For example, linear regression models assume a Gaussian distributed dataset. We mitigated this risk by applying a logarithmic transformation, a square root transformation, and a Yeo-Johnson[47] transformation to the skewed features with the aim of mutating the dataset into a new dataset that can be used with most ML algorithms. The logarithmic transformation shrinks a long heavy-tailed distribution of a feature X and consolidates its smaller values into larger ones. Therefore, it roughly transforms the data distribution into a normal distribution and reduces the effect of outliers. Likewise, we applied a square root transformation to all positive features to magnify their small numbers and to counterweight larger ones. However, it is not possible to apply the logarithmic or square root transformation to negative values; therefore, we applied a Yeo-Johnson Eq. (1) transformation to the negative skewed features.

$$y(\lambda)=\begin{cases}\dfrac{(y+1)^{\lambda}-1}{\lambda}, & \text{when } \lambda\neq 0,\ y\geq 0\\ \log(y+1), & \text{when } \lambda=0,\ y\geq 0\\ \dfrac{(1-y)^{2-\lambda}-1}{\lambda-2}, & \text{when } \lambda\neq 2,\ y<0\\ -\log(1-y), & \text{when } \lambda=2,\ y<0\end{cases} \tag{1}$$

Additionally, as suggested in Refs. 28 and 29, to minimize the influence of outliers and the inter-individual physiological variation in adapting to a stressor, we scaled the datasets by applying a scaler $S_c(X)$ to every data point X_i of each feature X in Eq. (2). $S_c(X)$ removes the feature's median and uses its 25th and 75th quantiles to readjust the data points.

$$S_c(X)=\frac{X_i-\text{median}(X)}{Q_3(X)-Q_1(X)} \tag{2}$$

This feature engineering resulted in as many as 94 features. It is possible that some of these features have correlations with others and that some are not very relevant to stress prediction. There might thus be a need to decrease the number of dataset attributes—not least because this will reduce the computational requirements of the resulting predictive models—but most importantly because it could increase the generalization of models. We computed the mean decrease in impurity (MDI) of each feature using Eq. (3), i.e., the mean loss in the impurity index of all trees of a random forest (RF) when that particular feature is used during tree splitting.

$$G_k=\sum_{k=1}^{k} p_k(1-p_k) \tag{3}$$

Here, k is the total number of features and p_k the proportion of a single HRV feature k. We ranked all the features and heuristically selected only the features with high MDIs and removed those with very small ones. Table 2 shows a summary of the resulting datasets. (The datasets are available at https://www.kaggle.com/qiriro/ieee-tac)

2.4 Stress prediction

We developed regression stress prediction models based on each participant's self-reported stress and mental load scores (in terms of the NASA-TLX and SSSQ for the SWELL and WESAD datasets, respectively), and based on the subtle changes in the participants' EDA and HRV signals. We also classified the stress based on the experimental conditions discussed in Sect. 2.1. We trained and evaluated three stress prediction models:

(1) Person-specific models—They were developed using RF models (Table 3). All person-specific models were trained and tested exclusively on the physiological samples of the same person and validated using 10-fold CV.

Table 2
Summary of the downsampled datasets.

	Signal	# of samples	# of features	# of classes
SWELL	HRV	204885	75	3
	EDA	51741	46	3
WESAD	HRV	81892	40	2
	EDA	20496	45	2

Table 3
Hyperparameters of the RF models.

Hyperparameters	Classification	Regression
Number of trees	1000	1000
Maximum depth of trees	2	2
Best split max features	$\sqrt{number\ of\ features}$	$\frac{1}{3}(number\ of\ features)$

Algorithm 1
Model calibration.

Input: machine learning algorithm h_m
Data:
- Samples $sample_{generic}$ collected from n persons
- Calibration samples $sample_{calibration}$ that belong to q unseen persons such that $q \ll n$

Output: trained calibrated model h_m'
/* mix the calibration samples and the generic samples */
$D' \leftarrow \varnothing$
$D' \leftarrow shuffle(sample_{generic} \cup sample_{calibration})$
/* train the model h_m on dataset D' */
$h_m' \leftarrow h_m(D')$
return h_m'

(2) Generic models—They were also developed using RF models (Table 3). We used LOSO-CV to assess how a generic model would perform in predicting the stress of unseen people (i.e., people whose samples were not part of the training set) as follows: In a dataset of n subjects, for each subject S_i, we trained the ML model on the data of $(n - 1)$ subjects and validated its performance on the left-out subject S_i.

(3) Hybrid calibrated models—As we expected (see discussion in Sects. 1 and 3), the generic models performed poorly compared with the person-specific models. To mitigate this discrepancy, we devised a hybrid technique that derives a personalized stress prediction model from samples collected from a large population. The technique (Algorithm 1) consists of incorporating a few person-specific samples (calibration samples) in a generic pool of physiological samples collected from a large group of people and training a new model from this heterogeneous data. In this paper, for a dataset with N subjects, we used the calibration algorithm with $q = 4$ and $n = N - q$, i.e., we reserved the physiological samples of four randomly selected subjects as "unseen subjects" and used the data of the remaining $n = N - q$ subjects as "generic samples". All calibration models were trained on extremely randomized trees models (ExtraTrees) whose key hyperparameters are summarized in Table 4.

Table 4
Hyperparameters of the ExtraTrees models.

Hyperparameters	Classification	Regression
Number of trees	1000	1000
Maximum depth of trees	16	16
Best split max features	$\sqrt{number\ of\ features}$	$\frac{1}{3}(number\ of\ features)$

3. Results and Discussion

3.1 Individual differences in stress prediction

All the person-specific models (i.e., the models that predict the stress of a preordained person) achieved an unrivaled performance. This high performance is, however, deceptive in that it would not generalize on yet unseen people. Indeed, the generic models (i.e., the models that predict the stress of any person) performed very poorly as shown in Figs. 1 and 2.

It is, of course, reasonable to assume the models over-fitted. However, there is no indication that this was the case. First, we validated all the person-specific models using 10-fold CV, and observed that 10-fold CV produced consistent predictions with a very low standard deviation between the 10-folds. *K*-fold CV provides an unbiased estimation of the performance of a model because it tests how well the *k* different parts of the training data perform on the model. Therefore, if the models had overfitted, they would have underperformed when tested on some folds. In our case, all folds achieved similar performance characteristics. [Interested readers are referred to the detailed tables in the supplementary materials (see Sect. 5 for more details)] Secondly, all the models use a very simple RF model (Table 3) that is unlikely to overfit. We believe that the models do not overfit because they consist of a large number of shallow trees (1000 trees, maximum depth = 2) and that each model has a small number of best split features. A small number of best split features allows the model to create more diverse and less correlated trees; therefore, the aggregation of the different trees results in a model with low generalization error variance and high stability.[48] Moreover, the trees are shallow (maximum depth = 2) to reduce the model's complexity and thus minimize overfitting. Finally, our results are similar to others in the literature: in general, person-specific models achieve accuracy greater than 90%,[15,23,24,27,31,35] while generic models always under-perform.[20,24,27]

The drop in accuracy, when tested on unseen subjects, is also not unusual, as already explained (Sect. 1). Indeed, the models cannot learn the inter-subject physiological differences in how people respond to stressors. To double-check this verdict, we added a subject id as a control prediction feature to the datasets. The subject id was used to monitor the subject to whom each sample in the datasets belongs to and to probe how much each model is influenced by knowing the origin of each sample. The influence of the subject id on the model is assessed by comparing the importance (in terms of the MDI) of the subject id with that of other attributes of the dataset. The MDI score of an attribute reveals how much the said attribute contributes to making the final prediction of a model. We found that, in all datasets, the subject id has the highest MDI and thus is the most critical attribute for stress prediction.

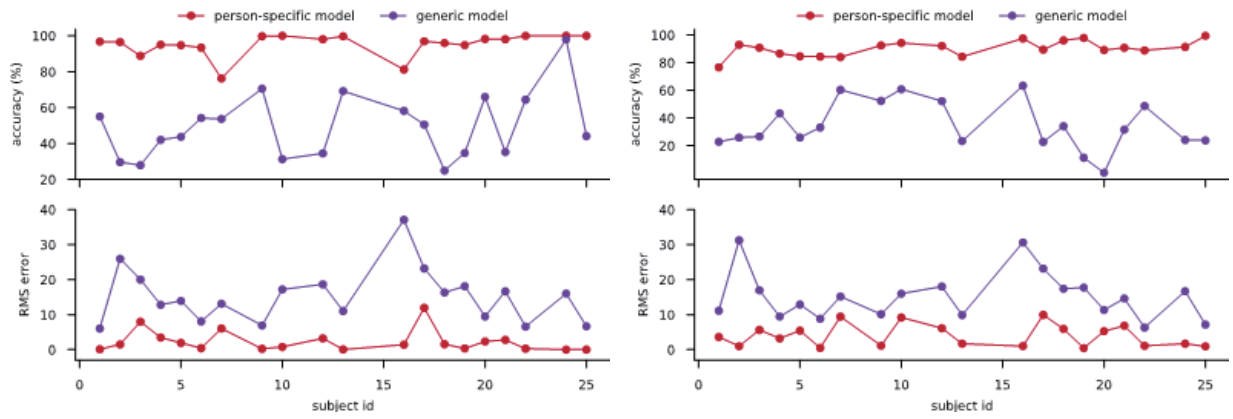


Fig. 1. (Color online) Performance comparison between person-specific and generic models trained on the SWELL datasets. For all subjects, the person-specific classification models (classification into three classes) achieved high accuracy, and the regression models (based on NASA-TLX (max = 55.5, min = 26.1, std = 14.8) have a small RMSE (e.g., 95.2 ± 0.5%, 2.3 ± 0.1 RMSE for the HRV dataset). However, because of the inter-individual differences in the response to stress, all the generic models performed poorly (e.g., 42.5 ± 19.9%, 15.3 ± 7.9 RMSE for the HRV signal), and there is a very large variation in performance among the subjects. (a) SWELL HRV and (b) SWELL EDA datasets.

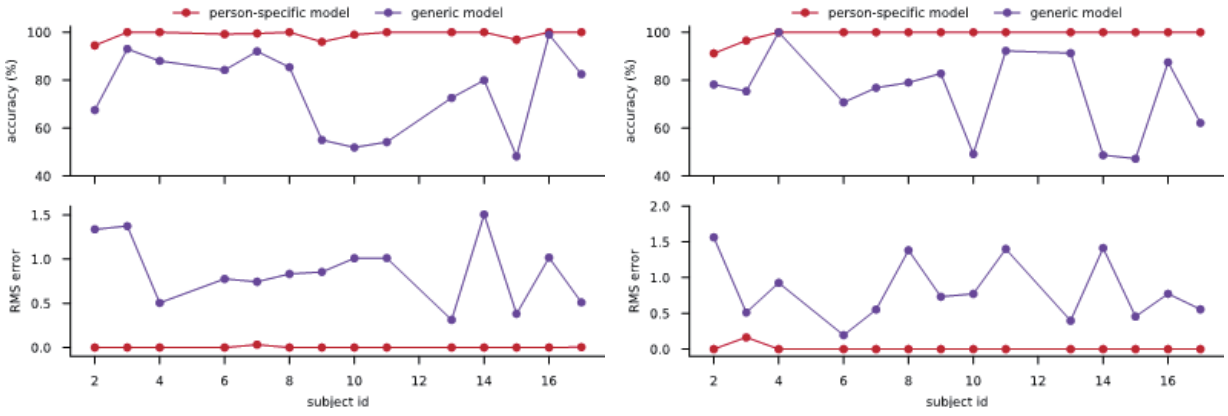


Fig. 2. (Color online) Performance comparison between person-specific and the generic models trained on the WESAD datasets. For all subjects, the person-specific classification models (classification into two classes) achieved high accuracy, and the regression models based on SSSQ (max = 3.9, min = 3.0, std = 0.8) have small RMSE (e.g., 98.9 ± 2.4%, 0.002 ± 0.001 RMSE for the HRV signal). However, because of the differences in how different subjects react to stress, all the generic models performed poorly, and there is a vast performance variation between the subjects (e.g., 83.9 ± 13.2%, 0.8 ± 0.3 RMSE for the HRV signal). Also note that, compared with the SWELL datasets (Fig. 1.), the classification models seemingly achieved a higher level of performance because the datasets contain only two classes. (a) WESAD HRV and (b) WESAD EDA datasets.

We evaluated the classification models by computing their accuracy, precision, recall, and F_1 score when tested on the test datasets. For the regression models, their performance is evaluated by calculating their MAE and root mean square error (RMSE).

Additionally, as shown in Figs. 1 and 2, unlike the person-specific models, because each subject has a unique response to stress, the generic model's performance varies widely among the different subjects. Accordingly, using generic stress prediction models would lead to unpredictability and low performance compared with using person-specific models. This discrepancy in performance highlights the far-reaching importance of inter-individual physiological differences that makes it hard for a generic stress prediction model to generalize to new unseen people. As already discussed by other researchers, one-size-fits-all stress prediction models cannot work well because people express stress differently.

Furthermore, there is a wide gap in how generic models perform on different subjects. This wide gap implies that, if a system uses a generic model for stress prediction, in practice, its prediction would seem virtually arbitrary, making it very laborious to troubleshoot when the system has bugs. Therefore, an effective system would need to rely on non-economically viable person-specific models.

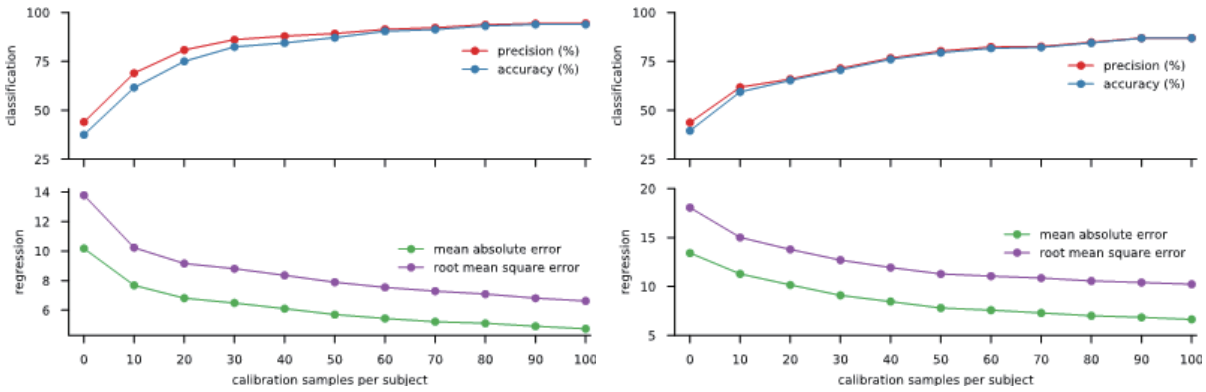
3.2 Generic stress model calibration

While it was possible to slightly increase the performance of the generic models (e.g., by using complex stacked models), it was clear that the performance of the person-specific models always greatly exceeded that of the person-independent models (Figs. 1 and 2). Furthermore, it was not possible to reliably optimize hyperparameters. Tuning the hyperparameters involves guesswork and is an erratic process given that the distribution of each subject is unique; for this reason, finding hyperparameters for a model that works well for all subjects is a futile endeavor.

In an attempt to improve the generalization of models on unseen people, we investigated how each model would perform if it knew little information about the previously unseen subjects. Consequently, we devised a technique that derives a personalized model from the data collected from a large group of people (see Algorithm 1). In this paper, we used half of the data from $q = 4$ randomly selected subjects as the calibration samples and the remaining half to test the performance of the calibrated models. The data of the remaining $n = N - q$ subjects were used as generic samples. In one sense, the calibration samples serve as "the fingerprints of a person," i.e., they encode the "uniqueness" of an individual using tiny physiological samples of that person.

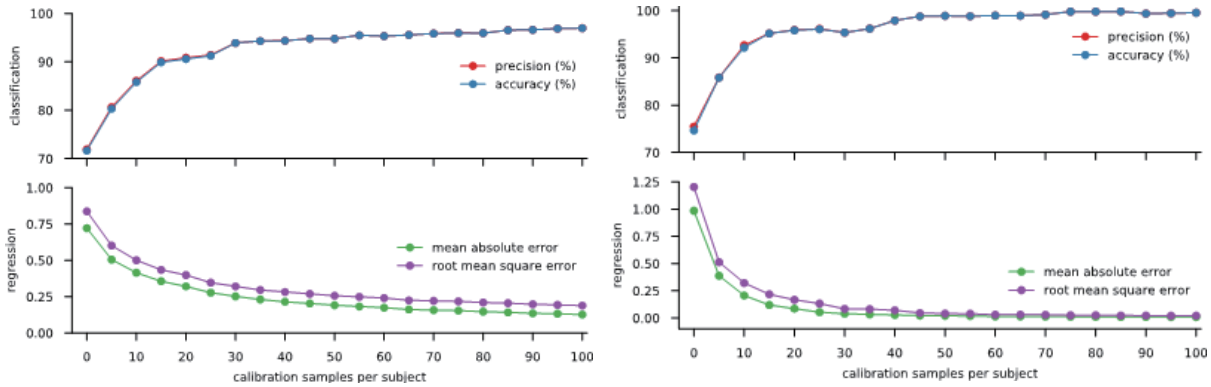
When we applied this technique to stress prediction on the two datasets, the performance of all the models significantly increased as follows, even when we only used a few calibration samples (see Figs. 3 and 4 for more details):

- The RMSE and MAE sharply decreased when we used a few calibration samples, and this was observed for both the model trained on the EDA datasets and the model trained on the HRV dataset. For instance, for the model trained on the HRV signal of the SWELL dataset, the MAE decreased from 10.1 to 7.6 when we only used 10 calibration samples per unseen subject. Likewise, this error dropped even further when we used 100 calibration samples (MAE = 4.7, RMSE = 6.6)
- In a similar manner, the performance of the classification models notably increased when we used a few calibration samples. For instance, for the model trained on the HRV signal of the SWELL dataset, the accuracy, precision, and recall respectively increased from 37.5, 44.0,



(a) (b)

Fig. 3. (Color online) Performance of the hybrid model trained on the SWELL datasets. Without the calibration samples, both the regression and classification models performed crudely. However, when a few person-specific calibration samples were used for calibration, their performance steadily improved. (a) SWELL HRV and (b) SWELL EDA datasets.



(a) (b)

Fig. 4. (Color online) Performance of the hybrid model trained on the WESAD datasets. Without the calibration samples, both the regression and classification models performed crudely. However, when a few person-specific calibration samples were used for calibration, their performance steadily improved. (a) WESAD HRV and (b) WESAD EDA datasets.

and 37.50% to 61.6, 69.0, and 61.6% when we used only 10 calibration samples per unseen subject and culminated in 93.9% accuracy, 94.4% precision, and 93.9% recall with 100 calibration samples per subject.

The increase in performance due to the few person-specific calibration samples highlights the influence of person-specific biometrics in predicting stress. In Ref. 38, the authors showed that, when inter-individual physiological differences are not accounted for, a stress predictive model may perform no better than a model with no learning capability. Our result confirms their findings. Nevertheless, all humans share a common hormonal response to stress,[53] albeit a person's unique factors such as gender,[40] genetics,[54] personality,[44] weight,[55] and coping

ability[39] determine how he/she reacts to stress. Previous researchers (e.g., Refs. 23, 50, and 51) have achieved notable improvements in generic stress prediction models by clustering the subjects based on their physiological or physical similarity. Their methods are, however, not practical for mass-marker stress monitoring products because they rely on heuristic clustering methods, and there is no authoritative subject clustering criterion. Our proposed method is simpler and much cheaper for real-world deployment (see discussion in Sect. 4) and performs much better than any previously proposed generic model improvement technique.

4. Stress Monitoring in Offices

The above results suggest that to design a real-world stress monitoring system, it would be beneficial to rethink the trade-off between spending effort on collecting data and training a highly performing but costly person-specific model, versus using a hybrid model derived from a mixture of a few person-specific physiological samples and the physiological samples collected from a large population. The latter approach is less expensive and more flexible for deployment, and delivers performance characteristics comparable to those of person-specific models.

The architecture and deployment of a stress monitoring system that uses this technique will undoubtedly involve many technical challenges that are beyond the scope of this study. We encourage the interested reader to examine[2,5] for an exhaustive overview of these challenges. One of the biggest challenges is perhaps how to collect the required physiological signals unobtrusively. Indeed, the system should not interfere with a person's routine. At the same time, it should record the physiological signals meticulously, accurately, and at a sufficient sampling frequency because the quality of the physiological data affects the performance of the stress prediction models.[49] These stringent requirements necessitate a compromise between conflicting requirements. For instance, while an HRV signal recorded using the chest leads is always of the highest quality, its recording would hinder the person's normal life. Alternatively, the HRV signal could be obtained using a lower quality but less invasive PPG signal recorded from the person's wrist; many wearable devices (e.g., smartwatches and fitness trackers) with built-in PPG sensors exist. For example, the Empatica E4 wristband (https://www.empatica.com/research/e4/) might serve for this purpose. The device boasts of a high-resolution EDA sensor with a strong steel electrode that can continuously record both the tonic and phasic changes in skin conductance. As discussed in a recent article,[50] the Empatica E4 wristband has sufficient accuracy to record HRV under seated rest, paced breathing, and recovery conditions. However, it is not very reliable when its wearer makes wrist movements.

Another challenge is how to deploy the stress prediction models. The recent reviews on stress recognition[1,2] unanimously concluded that due to the physiological difference in how people react to stress, a stress monitoring system should adapt to every individual's physiological needs. Simple and probably the most accurate approach is to deploy each person's stress prediction model as a web service [e.g., Representational State Transfer (REST) web service] that can be used to predict the person's stress. Regrettably, such an approach is daunting, time-consuming, and expensive because in, for example, an office environment, it

would require the collection, cleaning, and labeling of new data and the training of a new model for each office employee. Moreover, once deployed, the resulting stress monitoring system will unquestionably not perform as expected because its performance would deteriorate with time considering that a person's stress is dynamic and affected by many factors.[38] Consequently, with this approach, a real-world system will need to periodically start over and collect, label, and train new models for each user to prevent performance degradation.

As implied by the results of this paper (see Sect. 3.2), an alternative and cost-effective method would be to derive a high-performance model from a combination of generic samples collected from a large population and few person-specific calibration samples. It would also be beneficial to automate the entire process. As an illustration, after training and testing a generic stress prediction model, it may be possible to create an automatic self-updating stress prediction model pipeline, depicted in Fig. 5, as follows:

STEP I Calibration sample collection—Once the stress monitoring system is deployed, at the beginning (at this point, it uses only a generic model), it is essential that its users take several self-evaluation surveys under different working conditions to allow the collection of self-evaluation ground truths that reflect the broad ranges of stressors that its users will likely go through. At the same time, each user's physiological signals are recorded using an unobtrusive wearable device (e.g., an Empatica E4 wristband) and saved in a database. Once the system has collected enough calibration

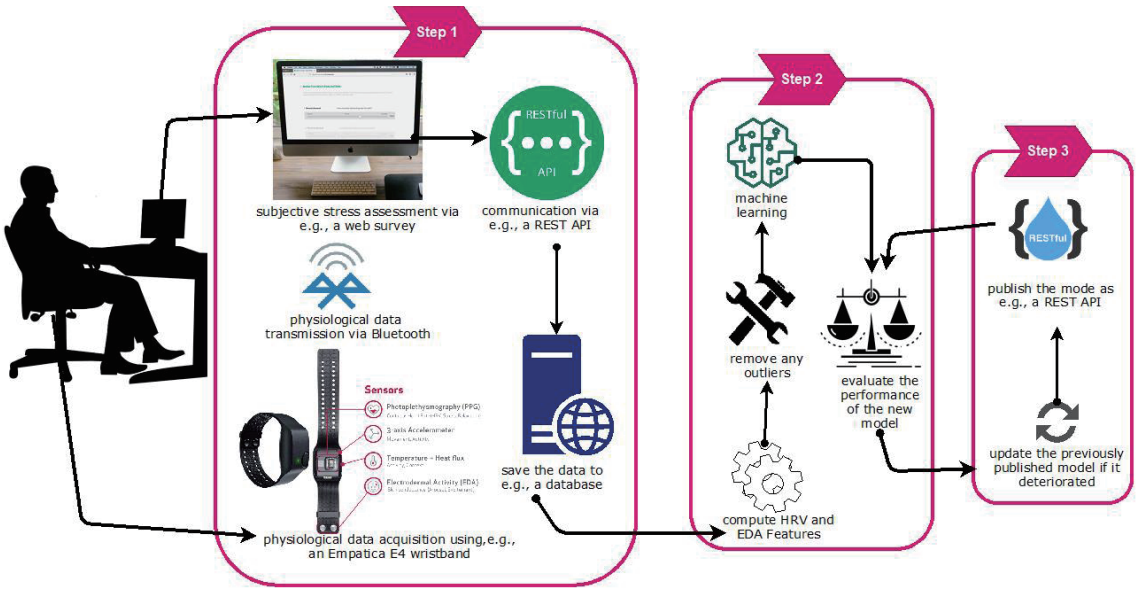


Fig. 5. (Color online) Simplified pipeline for a continuous stress monitoring model. Aeon's PPG and EDA signals are recorded using a wristband device. The signals are sent to a computing device where appropriate features (see Table 1) are computed, preprocessed (e.g., data cleaning and rebalancing), and sent to a remote server where they are used to predict the person's stress. For calibration purposes, the person also periodically provides self-assessment of his/her stress (e.g., via a web survey after the completion of work). This feedback is used to train a personalized stress prediction model, which is published and used as a RESTful API. When the model deteriorates, it is automatically updated using the periodic self-evaluations the system received from its users.

samples from the users, it automatically creates each user's personalized model by training a new model on a combination of the new user-specific data with the data used to train the generic model as shown in Algorithm 1.

STEP II Continuous ML—After these personalized models are deployed, the system periodically reminds its users to provide additional calibration samples by taking short self-report surveys periodically (e.g., via a web survey every time he/she finishes a major task of work) to give more feedback data to improve the user's personalized model. Indeed, with time, the models will be prone to the effect of concept drift, i.e., they will become invalid because their input data unpredictably change over time. In stress prediction, model drift is particularly inevitable because stress is inherently dynamic. The models thus need to adapt to changes. For example, when the system has received a specific number of new calibration samples from a user, it automatically tests their accuracy against the existing model. If this prediction indicates a deterioration of the model, the system will need to update the model to reverse the drift. There are many ways to achieve this. One approach would be to train a new model on a combination of the data of the generic model and the new calibration samples. This approach would, however, be computationally expensive and require significant time to retrain each user's model. Depending on the system, it may instead be more appropriate to incrementally train the existing model as the new data are received. This approach is faster because it does not require retraining the whole model when new data come in. Instead, this approach extends the existing model by, for example, combining the new data with a subset of the old data. Nevertheless, it is important to note that many ML algorithms do not support incremental learning and that, unless there is rigorous monitoring of the system, incremental learning may introduce predicaments such as strong drift and recurrence.

STEP III Calibrated model deployment—The model is published as, for example, a REST application program interface (REST API) and periodically updated depending on its performance as discussed in STEP II above.

Although there is a need to validate our assumptions, we believe that developing a continuous stress monitoring system based on this strategy would present the following benefits over existing approaches:

- Lower cost—For practicality, the existing approaches would require collecting and labeling the training data for each user. This process is costly and would incur high installation, support, and maintenance costs. Our approach would likely be less expensive because there would be no need to collect large quantities of new data from each user. Instead, only a few user-specific samples would be required.
- Practicality—All high-accuracy stress prediction methods rely on person-specific models. As already discussed, this approach is suboptimal when applied to new unseen people. The alternative is to create person-specific models. While this approach performs excellently in predicting stress, it is not practical in real-world settings because it is not scalable to many users and would be very costly to implement, and, most importantly, it is not flexible to the expected dynamic changes in each user's stress. The proposed approach achieves stress

prediction accuracy comparable to that achieved by subject-dependent models while yet presenting enticing large-scale deployment benefits.

- Straightforward deployment—Once deployed, each user's person-specific model can be generated using a very small number of user-specific samples that can be unobtrusively collected using, for example, the approach proposed in Ref. 71, in which each user can self-evaluate (in terms of NASA-TLX and SSSQ) his/her stress level via a smartphone application. The self-evaluation would serve as a person-specific calibration with the generic model. Over time, when the model degrades due to the person's stress dynamics, a few new physiological samples would be collected and used to train and update each person's model periodically.

Although the results of this study are encouraging, this system still has many limitations. Notably, the study did not validate the proposed approach in real-world settings, and it reached its conclusion using only two datasets with a small homogeneous group of subjects. Furthermore, designing a continuous stress monitoring system using the proposed approach requires extraordinary care because external factors can influence both the EDA and the HRV. In particular, the EDA signal, while often heralded as one of the best indicators of stress,[3] has significant drawbacks. The EDA is a result of electrical changes that occur when the skin receives signals from the nervous system. Under stress, the skin's conductance changes due to a subtle increase in sweat that leads to a decrease in the skin's electrical resistance. The variation in skin conductivity is, however, influenced by other unrelated factors such as the person's hydration, ambient temperature, and ambient humidity. Moreover, for the same person, an EDA signal may fluctuate from one day to another. Additionally, because stress is intrinsically multifaceted (it consists of physiological, behavioral, and affective responses), it is imperative to take into consideration its context (i.e., where, what, when, who, why, and how). This approach may yield better and more predictable results even when tested under real-life conditions.

It is also important to highlight that the deployment of a stress monitoring system based on our approach still poses technical and cost challenges. The system would require considerable upfront investments and would be undoubtedly beyond the budget of a small business. However, the investment might be rewording for a large business. In our previous studies, we showed that it is possible to predict the thermal comfort of people using the variations in their HRV[53] and highlighted the energy-saving potential of this approach.[54] Therefore, the positive spillovers that might result from using the system may outstrip the initial investment because, in a responsive smart office, the system can be used as part of a multipurpose system that uses the physiological signals of office occupants for preventive medicine and stress management while efficiently providing thermal comfort at low energy. Additionally, there are enabling technologies that would make these challenges easier. For example, IBM's Watson Studio (https://www.ibm.com/cloud/machine-learning) offers tools that simplify the development and deployment of predictive models. In our proposed stress monitoring system, Watson Studio, which requires little or no programming experience, could be used to automate steps 1 and 2 (see Fig. 5) including model deterioration monitoring and deployment as a REST API.

5. Conclusion

Despite the extensive literature on stress recognition and the potential economic and health benefits of stress monitoring, there is not yet a robust real-world stress recognition system. The most reliable and rigorous methods use a fusion of multimodal signals [e.g., physiological (such as HRV, EDA, EEG, EMG, skin temperature, respiration, pupil diameter, and eye gaze), behavioral (keystrokes and mouse dynamics, and sitting posture), facial expression, speech, and mobile phone use patterns]. This approach, however, raises both practical challenges (e.g., real-time multimodal data acquisition, data fusion, and data integration) and user privacy concerns (e.g., the implications of recording a person's computer keystrokes, video, and speech), and is not feasible in real-world settings because of company-wide computer security policies or international workplace privacy laws.

In contrast, most practical stress monitoring methods that use physiological signals are idiosyncratic because stress is inherently subjective and felt differently by each person. Therefore, methods that use an ML model that uses physiological signals fail to generalize well when predicting the stress of new unseen people. Thus, they are not suitable for a real-world stress monitoring system. Only person-specific models are accurate enough for this task. Unfortunately, unlike the generic models, person-specific models are inflexible and costly to deploy in real-world settings because they require the collection of new data and the training of a new model for every user of the system. In an office environment, this entails spending precious resources. Moreover, because stress is inherently dynamic, these models will need expensive periodic updates to collect and retrain them to prevent the system from deterioration due to concept drift.

In this paper, we proposed a cost-effective hybrid stress prediction approach. Our method is based on the fact that humans share similar hormonal responses to stress. However, every person possesses unique factors (e.g., gender, age, weight, and coping ability) that differentiate the person from others. Therefore, we hypothesized that it may be possible to improve the generalization performance of a generic stress prediction model trained on a large population by deriving a personalized model from a combination of samples collected from a large group of people with a few person-specific samples. In a sense, the calibration samples serve as the "fingerprint" of a person and they introduce his/her "uniqueness" into the new model.

We tested our method on two stress datasets and found that our approach performed much better than the generic models. Furthermore, we surmised that to create a practical stress monitoring system, this approach would be cost-effective and practical to deploy in real-world settings, and we discussed some of its technical limitations.

References

1 D. Carneiro, P. Novais, J. C. Augusto, and N. Payne: IEEE Trans. Affect Comput. **10** (2019) 237.
2 P. Schmidt, A. Reiss, R. Dürichen, and K. V. Laerhoven: arXiv CoRR (2018). http://arxiv.org/abs/1811.08854
3 A. Alberdi, A. Aztiria, and A. Basarab: J. Biomed. Inf. **59** (2016) 49.
4 E. D. Kirby, S. E. Muroy, W. G. Sun, D. Covarrubias, M. J. Leong, L. A. Barchas, and D. Kaufer: Elife **2013** (2013) 1.
5 F. S. Dhabhar, W. B. Malarkey, E. Neri, and B. S. McEwen: Psychoneuroendocrinology **37** (2012) 1345.
6 C. Tennant: J.Psychosom. Res. **51** (2001) 697.

7 T. W. Colligan and E. M. Higgins: J. Workplace Behav. Health **21** (2005) 89.
8 EU-OSHA, Psychosocial risks and stress at work - safety and health at work, 2017.
9 J. M. Peake, G. Kerr, and J. P. Sullivan: Front. Physiol. **9** (2018) 1.
10 A. H. Marques, M. N. Silverman, and E. M. Sternberg: Neuroimmunomodulation **17** (2010) 205.
11 H. Hellhammer and C. Kirschbaum: Psychoneuroendocrinology **19** (1994) 313.
12 K. P. Eisen, G. J. Allen, M. Bollash, and L. S. Pescatello: Comput. Human Behav. **24** (2008) 486.
13 S. Järvelin-Pasanen, S. Sinikallio, and M. P. Tarvainen: Ind. Health **56** (2018) 500.
14 P. Adams, M. Rabbi, T. Rahman, M. Matthews, A. Voida, G. Gay, T. Choudhury, and S. Voida: Proc. 8th Int. Conf. Pervas. Comput. Paradig. Ment. Heal. (2014) 72.
15 P. Melillo, M. Bracale, and L. Pecchia: Biomed. Eng. Online **10** (2011) 96.
16 B. Cinaz, B. Arnrich, R. La Marca, and G. Tröster: Pers. Ubiquitous Comput. **17** (2013) 229.
17 K. S. Rahnuma, A. Wahab, N. Kamaruddin, and H. Majid: Proc. Int. Symp. Consum. Electron. (2011) 592.
18 C. Z. Wei: Adv. Mater. Res. **709** (2013) 827.
19 S. Poria, E. Cambria, R. Bajpai, and A. Hussain: Inf. Fusion **37** (2017) 98.
20 S. Koldijk, M. A. Neerincx, and W. Kraaij: IEEE Trans. Affective Comput. **9** (2018) 227.
21 O. M. Mozos, V. Sandulescu, S. Andrews, D. Ellis, N. Bellotto, R. Dobrescu, and J. M. Ferrandez: Int. J. Neural Syst. **27** (2017) 1650041.
22 M. Gjoreski, H. Gjoreski, M. Lûstrek, and M. Gams: Proc. 2016 ACM Int. Conf. Pervas. Ubiq. Comput. (2016) 1185.
23 J. A. Healey and R. W. Picard: IEEE Trans. Intell Transp. Syst. **6** (2005) 156.
24 Y. Nakashima, J. Kim, S. Flutura, A. Seiderer, and E. André: Pervas. Comput. Paradig. Ment. Heal. **604** (2016) 23.
25 A. Alberdi, A. Aztiria, A. Basarab, and D. J. Cook: Int. J. Ind. Ergon. **67** (2018) 13.
26 P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laer-hoven: Proc. 2018 Int. Conf. Multimodal Interaction (2018) 400.
27 J. Kim and E. Andre: IEEE Trans. Pattern Anal. Mach. Intell. **30** (2008) 2067.
28 B. Lamichhane, U. Großekathöfer, G. Schiavone, and P. Casale: Lect. Notes Inst. Comput. Sci. Soc. Telecommun. Eng. LNICST **181** (2017) 259.
29 L. Kogler, V. I. M¨uller, A. Chang, S. B. Eickhoff, P. T. Fox, R. C. Gur, and B. Derntl: Neuroimage **119** (2015) 235.
30 J. Wang, M. Korczykowski, H. Rao, Y. Fan, J. Pluta, R. C. Gur, B. S.McEwen, and J. A. Detre: Soc. Cogn. Affective Neurosci. **2** (2007) 227.
31 A. Liapis, C. Katsanos, D. Sotiropoulos, M. Xenos, and N. Karousos: ACM Int. Conf. Proc. Ser. **1** (2015) 323.
32 J. Hernandez, R. R. Morris, and R. W. Picard: Lect. Notes Comput. Sci. **6974** (2011) 125.
33 D. M. Almeida, J. R. Piazza, and R. S. Stawski: Psychol. Aging **24** (2009) 819.
34 N. Attaran, A. Puranik, J. Brooks, and T. Mohsenin: IEEE Trans. Circuits Syst. II Express Briefs **65** (2018) 2032.
35 J. Aigrain: Ph.D. Thesis, Université Pierre et Marie Curie (2016).
36 Q. Xu, T. L. Nwe, and C. Guan: IEEE J. Biomed. Heal. Inf. **19** (2015) 275.
37 N. Schneiderman, G. Ironson, and S. D. Siegel: October **1** (2008) 1.
38 E. Charmandari, C. Tsigos, and G. Chrousos: Annu. Rev. Physiol. **67** (2005) 259.
39 S. Wüst, I. S. Federenko, E. F. C. van Rossum, J. W. Koper, R. Kum-sta, S. Entringer, and D. H. Hellhammer, Ann. N. Y. Acad. Sci. **1032** (2004) 52.
40 E. Childs, T. L. White, and H. de Wit: Behav. Pharmacol. **25** (2014) 1.
41 S. U. Jayasinghe, S. J. Torres, C. A. Nowson, A. J. Tilbrook, and A. I. Turner: Endocr. Connect. **3** (2014) 110.
42 S. Koldijk, M. Sappelli, S. Verberne, M. A. Neerincx, and W. Kraaij: Proc. 16th Int. Conf. Multimodal Interaction - ICMI '**14** (2014) 291.
43 S. G. Hart and L. E. Staveland: Adv. Psychol. **52** (1988) 139.
44 C. Kirschbaum, K. M. Pirke, and D. H. Hellhammer: Neuropsychobiology **28** (1993) 76.
45 W. S. Helton and K. Näswall: Eur. J. Psychol. Assess. **31** (2015) 20.
46 M. Malik, J. T. Bigger, A. J. Camm, R. E. Kleiger, A. Malliani, A. J. Moss, and P. J. Schwartz: Eur. Heart J. **17** (1996) 354.
47 I.-K. Yeo: Biometrika **87** (2000) 954.
48 P. Probst, M. N. Wright, and A. Boulesteix: Wiley Interdiscip. Rev. DataMin. Knowl. Discovery **9** (2019) e1301.
49 A. Chowdhury, R. Shankaran, M. Kavakli, and M. M. Haque: IEEE Sens. J. **18** (2018) 3055.
50 L. Menghini, E. Gianfranchi, N. Cellini, E. Patron, M. Tagliabue,and M. Sarlo: Psychophysiology **56** (2019) e13441.

51 N. E. Bush, G. Ouellette, and J. Kinn: Mil. Med. **179** (2014) 1453.
52 J. Bakker, M. Pechenizkiy, and N. Sidorova: 2011 Int. Conf. Data Mining Workshops (2011) 573.
53 K. N. Nkurikiyeyezu, Y. Suzuki, and G. F. Lopez: J. AmbientIntell. Humaniz. Comput. **9** (2018) 1465.
54 K. Nkurikiyeyezu, Y. Suzuki, P. Maret, G. Lopez, and K. Itao: SICE J. Control. Meas. Syst. Integr. **11** (2018) 312.
55 S. U. Jayasinghe, S. J. Torres, C. A. Nowson, A. J. Tilbrook, A. I. Turner: Endocr. Connect. **3** (2014) 110. https://doi.org/10.1530/EC-14-0042

About the Authors

Kizito Nkurikiyeyezu is a Ph.D. candidate at Aoyama Gakuin University, Japan. He received his B.Sc. in electrical engineering and M.Sc. in electrical and computer engineering from Oklahoma Christian University, Oklahoma City, U.S. His doctoral research takes a multidisciplinary approach and investigates the possibility of estimating a person's thermal comfort level from the fluctuations in his/her physiological signals and using appropriate constrained optimization algorithms to provide an optimum and personalized thermal comfort using the least possible energy. (kizito@wil-aoyama.jp)

Anna Yokokubo received her M.S. in human-computer interaction (HCI) in 2012 from Ochanomizu University, Tokyo, Japan. From 2012 through 2017, she worked at Canon Inc., where she contributed to the development of healthcare technology. She is currently working as a research assistant at Aoyama Gakuin University, Kanagawa, Japan. Her research interests are in human-computer interaction, information design, and user experience design. (yokokubo@it.aoyama.ac.jp)

Guillaume Lopez received his M.E. in computer engineering from INSA Lyon, France, and his M.Sc. and Ph.D. in environmental studies from the University of Tokyo, Japan in 2000, 2002, and 2005, respectively. From 2005, he worked as a research engineer at Nissan Motor Corp. and as a project dedicated assistant professor at the University of Tokyo from March 2009. In April 2013, he joined Aoyama Gakuin University as an associate professor of the Department of Integrated Information Technology. His research interests include lifestyle enhancement and healthcare support based on intelligent information systems using wearable sensing technology. His professional memberships include the ACM, AHI, IEEE, IPSJ, and SICE.
(guillaume@it.aoyama.ac.jp)

Supplementary Material

Additional supporting information is available online (Available at https://www.kaggle.com/qiriro/sm-paper) in our public repository. The repository contains the following more detailed information and the source code to replicate our findings:

- Source code we developed for this research
- Dataset of the computed HRV and EDA features
- HRV and EDA feature importance with and without subject id added to the datasets (see Sect. 2.3)
- Tables of the performance of the person-specific and generic models (refer to Sect. 2.3)
- Tables of the performance of the calibrated models (see details in Sect. 2.3)