The Influence of Person-Specific Biometrics in Improving Generic Stress Predictive Models

Kizito Nkurikiyeyezu, Anna Yokokubo, and Guillaume Lopez
Wearable Information Lab
Aoyama Gakuin University, Japan

Abstract—Because stress is subjective and is expressed differently from one person to another, generic, one-size-fits-all stress prediction machine learning models perform crudely. Only person-specific models yield reliable results, but they are not flexible and costly to deploy in real-world. For illustration, in an office environment, a stress monitoring system that uses person-specific models would require collecting new data and training 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 a person-specific-like stress prediction model from physiological samples collected from a large population. We validate our approach on two stress datasets. The result shows that our technique significantly enhances the performance of the generic model. For instance, the accuracy increased from 57.7% to 73.9% when we only used 10 person-specific calibration samples. We also introduce a blueprint for a continuous stress monitoring system based on our strategy, and we debate its merits and limitation.

Index Terms—continuous stress monitoring, physiological computing, heart rate variability, electrodermal activity, smart buildings

1 Introduction

Occupational stress is well-researched [1] [2] [3] [4] [5], though not least 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. Admittedly, although a small amount of stress is benign and even auspicious because it provides the necessary gumption to survive the tribulations of the modern workplace [6] [7], chronic stress (i.e., enduring stress) has detrimental repercussions. Physiological and psychological disorders [8] [9], 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-out workers have suboptimal productivity, are prone to higher job absenteeism and presenteeism, and are more predisposed to sickness [9], [11].

The importance of overcoming stress at work is primordial to the well-being of the workers and the bottom line of any business. Nevertheless, at the moment, there exists no mainstream real-world stress monitoring method [12]. The most reliable stress monitoring strategies rely on directly measuring the level of the stress-inducing hormones (e.g., salivary and cortisol concentration in sweat [13] [14]) and on psychological evaluations performed by psychologists. However, these procedures are neither suitable nor practicable for continuously monitoring stress in the workplace settings because they are obtrusiveness and are carried out sporadically. Moreover, in the case of physiological evaluations, people are reluctant to reveal their work stress honestly [15]. Luckily, stress spawns detectable physiological, psychological, and behavioral changes that can be used for automatic early stress recognition in office environments [1] [5]. For example, acute occupational stress decreases a person's Heart Rate Variability (HRV) and his parasympathetic activation [16]. Besides, there is substantial research that indicates that it is possible to indirectly monitor stress using physiological signals such as the Electrodermal Activity (EDA) [17], the HRV [18] [19], the Electroencephalogram (EEG) [20], and the Electromyography (EMG) [21].

Although there is a surfeit of publications [1] [3] [22] [5] on automatic stress prediction, at the moment, apart from a few niche circumstances and non-scientifically proven consumer products, there exist no effective real-world system that automatically and unobtrusively monitor people's stress [12]. On the one hand, some of the proposed methods (e.g., EEG based stress monitoring) are outright impractical because they are too obtrusive. On the other hand, the most accurate approaches (e.g., [23], [22] and [24]) predict stress using a fusion of multiple sensors data (e.g., audio, video, computer logging, posture, facial expression, and physiological features). These methods, however, raises privacy and security concerns (e.g., the implication of user's computer keystrokes logging, video recording, and speech recording), and, cannot be used in the real-world settings because of company-wide computer security policies or due to international workplace privacy regulations. Finally, the most practical and unobtrusive stress monitoring methods (e.g., [25] [26] [27] [28])—which are mostly based on physiological signal that are recordable on people's wrist (e.g., Photoplethysmography (PPG) and EDA) —are not yet mainstream to the general consumers despite their potential economic and health benefit. The lack of a viable stress monitoring products, despite the extensive research on occupational stress, the availability of enabling technology (e.g., smartphones with on-wrist HRV and EDA sensors) and despite the immense economical 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 [3] scrutinized the published literature and noted the striking

corresponding author: e-mail —kizito@wil-aoyama.jp

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 a any person). The article underscores that person-specific ML models (e.g., [29], [28], [30], [18], [31] and [32]) achieved an excellent prediction accuracy. Nevertheless, their predictions are person-specific—that is, the ML models would not generalize well in predicting stress of unseen people; therefore, cannot be used in creating mass-market stress monitoring products. On the contrary, the pragmatic generic and person-independent solutions (e.g., [33], [23], [34], [25], [35], and [36]) generally have a much lower stress prediction accuracy; accordingly, they are equally a poor choice for creating mass-market stress monitoring devices. For example [36] achieved a 95.0 % emotion recognition accuracy using person-specific ML models; however, the same approach resulted in a mere 70% accuracy when applied to a person-independent classification model. In a like manner,the authors in [37] 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. Their person-specific ML models achieved a 97% but the generic ones dwindled to a mere 42% accuracy. Their results resemble that in [23], which achieved a 90% accuracy when using a person-specific stress classification models. However, when applied the same approach to predict the stress of new subjects, its performance ebbed to a meager 58.8±11.6 % accuracy.

These mediocre outcomes are expected. First stress is intrinsically idiosyncratic and depends on a person's uniqueness (e.g., his genetics) and his coping ability [38]. Second, there is credible evidence that there exist gender difference in how people respond to stress [39] and that men and women have a different feeling about stress because women tend to express a higher level of stress on self-report questionnaires [40] [41]. Third, a stressor that produces stress in one person will not necessarily trigger the same stress response in a different person [42] [43] [44] [45]. Finally, for the same person, there exist significant day-to-day variability in the cortisol awakening response, which may affect how that person responds to stress [46]. As a result, a practical stress monitoring scheme needs to take into account inter-individual and intra-individual differences, people's gender and the temporal variability of human stress among many other factors that influences how humans react to stress. The state of the art stress monitoring strategies (e.g., [47]) are person-specific. Unfortunately, this method is not realistic for creating a real-world product. A system based on 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.

Recent research has proposed diverse methods 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 the stress classes [48] [49]. The normalization improves the performance of the generic model but always underperforms compared to the person-specific ones. Furthermore, as [49, Chapter 5] noted, 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 [23] [50] [51]. These techniques are important contributions to producing an effective stress monitoring system. However, they also perform inadequately compared to 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 [50] and, given that many factors influence a person's stress [52], it is not clear what are the criteria for similarity to create the clusters.

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 samples collected from a large group of people. The proposed method hinges on the premise that all humans share a common hormonal response to stress [53], but that a person's unique factors such as gender [39], genetics [54], personality [43], weight [55] and his coping ability [38] differentiate how the person reacts to stress. Hence, we hypothesize that it could be possible to reuse a generic stress prediction model trained on a large group of people as a starting point for a personalized and more effective model. To confirm these assumptions, we tested this strategy on two major stress datasets. Our results showed a substantial improvement in the stress prediction models' performance even when we used only 10 calibration samples. In summary, in this paper:

(i) For all n subjects in the datasets, we trained and validated n person-specific regression and classification stress prediction models using a 10-fold cross-validation approach. The result shows, for all subjects, the classification models achieved a 100% classification accuracy and that the regression models had a near-zero mean absolute error (MAE).

(ii) We also used a Leave-One-Subject-Out Cross-Validation (LOSO-CV) to assess the performance of generic stress prediction models. For each subject, we trained a model on the data of *(n-1)* subjects and validated its performance on the left-out subject and repeated the process for all n subjects. The result shows that all models performed poorly ($accuracy = 56.0{\pm}4.7\%$) compared to person-specific models and that there was a wide performance variation between the subjects. 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.

(iii) We devised 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, e.g., office settings.

The rest of the paper highlights the importance of inter-individual differences in stress prediction and discusses the proposed methods. It also uses a typical workplace as a case study to discuss the advantages and limitations of using the proposed approach for a continuous stress monitoring.

2 METHODS

2.1 Stress datasets

We used two stress datasets to conduct this study. The first dataset —the SWELL dataset [58] —was collected at the Radboud University. This dataset is a result of experiments conducted on 25 subjects doing office work (for example writing reports, making presentations, reading e-mail and searching for information) who were exposed to quintessential 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, their facial expressions, their body postures, their electrocardiogram (ECG) signal, and their electrodermal activity (EDA) signal. The participants went through three different working conditions:

1) *no stress* —the participants performed the assigned tasks for a maximum of 45 minutes.
2) *time pressure* —each participant's time to finish the task was reduced to two-thirds of the duration that he/she took in 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 immaterial, and the participants did not need to take any action.

At the end of each experiment condition, each participant's perceived stress was assessed using a variety of self-report questionnaires, including the NASA Task Load Index (NASA-TLX) [59]. In this study, we focus on the NASA-TLX because it indicates a person's mental load based on a weighted average of multi-dimensional rating (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 [33]—was collected by researchers from the Robert Bosch GmbH and the University of Siegen in Germany. The dataset includes physiological (EDA, ECG, EMG, respiration signal and skin temperature) and acceleration signal that the researchers collected from 15 subjects to whom they exposed to three affective stimuli as follows:

1) *baseline condition*—the baseline condition aimed at generating a neutral affective state onto the participants and lasted for 20 minutes.
2) *amusement condition* —the subjects watched funny video clips. Each video clip is followed by a brief (5 seconds) of neutral condition. The amusement condition lasted 392 seconds.
3) stress conditions —the participants were subjected to the Trier Social Stress Test (TSST) [60] and asked to give a five-minutes public speech and to count down from 2023 by 17. If the subject made an error, he/she is requested to start over.

The amusement and the stress conditions were each followed by a meditation period to "de-excite" the participants back to the baseline conditions. Throughout the experiment, the participants provided five self-reports, including the Short Stress State Questionnaire (SSSQ) [61] 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 [56]. Each HRV feature (Table I) was computed on a five-minutes moving window as follows: first, we extracted an Inter-Beat Interval (IBI) signal from the peaks of the Electrocardiogram (ECG) signal of each subject. Then, we computed each HRV index on a 5-minutes IBI array. Finally, a new IBI sample is appended to the IBI array while the oldest IBI sample is removed from the beginning of the IBI array. The new resulting IBI array is 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 4Hz fourth-order Butterworth low pass filter and then smoothed with a moving average filter. Next, we computed the EDA features (Table I) on 10-minute moving window signal extracted from various EDA attributes (indices of SCR pulse onsets, SC peaks location and the SCR amplitudes) of the skin conductance response (SCR).

All the resulting datasets—especially the WESAD datasets—are inherently unbalanced because their experimental protocols dictated different duration. We downsampled the datasets by randomly discarded some samples from the majority classes to make the dataset balanced; therefore, to prevent the majority classes from overshadowing the minority classes. Furthermore, for the WESAD dataset, we altogether removed all sample corresponding to *amusement condition* because it is almost as short as the sliding window we would use for computing the feature.

2.3 Feature engineering

An inspection of histogram plots of the features computed in section 2.2 revealed that most features' data distribution is skewed. While this may not be an issue for some machine learning algorithms, in other cases, the distribution of the features is critical. For example, linear regression models expected a Gaussian distributed dataset. We mitigated this risk by applying a logarithmic transformation, a square root transformation, and a Yeo and Johnson [62] transformations to the skewed features. The application of the three transformations aimed to mutate the dataset into a new dataset that can be used with most machine learning algorithms. The logarithmic transform shrinks long heavy-tailed distribution of a feature X and bolsters 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 transform on all positive feature to magnify the features' small numbers and to counterweight larger ones. However, it not possible to apply neither the logarithm transformation nor the square root transform to negative values; therefore,

TAB. I. Selected heart rate variability (HRV) and electrodermal activity (EDA) features

HRV Features	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 interval of differences between adjacent RR intervals	
	SDRR_RMSSD	Ratio of SDRR over RMSSD	
	pNNx	Percentage of number of adjacent RR intervals differing by more than 25 and 50 ms	ref. [56]
	SD1, SD2	Short and long-term poincare plot descriptor of the heart rate variability	
	RELATIVE_RR	Time domain features(e.g., mean, median, SDRR, RMSSD) of the relative RR	see note a
	VLF, LF, HF	Very low (VLF), Low (LF), High (HF) frequency band in the HRV power spectrum	
	LF/HF	Ration of low (LF) and high(HF) HRV frequencies	
EDA Features	Time domain	Mean, max, min, range, kurtosis, skewness of the SCR	
	Derivatives	Mean and standard deviation of the 1st and second derivative of the SCR	
	Peaks	Mean, max, min, standard deviation of the peaks	
	Onset	Mean, max, min, standard deviation of the onsets	ref. [57]
	ALSC	Arc length of the SCR	see note b
	INSC	Inegral of the SCR	see note c
	APSC	Normalized average power of the SCR	see note d
	RMSC	Normalized room mean square of the SCR	see note e

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

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

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

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

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

we used a Yeo and Johnson (Eq. (1)) transformation to the negative skewed features.

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

Additionally, as suggested in [49], 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 $S(X)$ to every data point X_i of each feature X(Eq. (2)). $S(X)$ subtracts feature's media and uses its 25^{th} and 75^{th} quantiles to re-adjust the data points.

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

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

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

Where K is the total number of features and p_k the proportion of a single HRV feature k.We ranked the all the features and heuristically selected only the features with high MDI and removed those with very small ones. Table II summarizes the resulting datasets [1].

TAB. II. Summary of the downsampled datasets

	signal	# of samples	# of features	# of classes
SWELL	HRV	204885	57	3
	EDA	51741	36	3
WESAD	HRV	81892	41	2
	EDA	20496	48	2

TAB. III. Hyperparameters of the extremely randomized trees models

Hyperparameters	Classification	Regression
number of trees	500	500
maximum depth of the trees	16	16
best split max features	$\sqrt{\text{number of features}}$	$\frac{1}{3}(\text{number of features})$
bootstraped	Yes	Yes

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 experiment conditions

1. The datasets are available online in supplementary material

discussed in Section 2.1 All models are based on Extremely randomized trees (ExtraTrees) models whose key hyper-parameters are summarized in Table III. We trained and evaluated the following three stress prediction models:

- *person-specific model*—we developed person-specific models by training and testing the models exclusively on the physiological samples of the same person. We validated each person-specific model using a 10-folds cross-validation.
- *generic model*—we used a Leave-One-Subject-Out Cross-Validation (LOSO-CV) to assess how a generic model would perform in predicting the stress of unseen people, (i.e., the people whose samples were not part of the training set): for each subject, we trained an Extremely randomized trees (ExtraTrees) model on the data of *(n-1)* subjects and validated its performance on the left-out subject.
- *hybrid calibrated model*—as we expected (see discussion in Section 1 and Section 3), the generic models performed poorly compared to 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 (the calibration samples) in a generic pool of physiological samples collected from a large group of people and to train 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 data of the remaining $n = N - q$ subjects as the "generic samples".

Algorithm 1: GENERIC 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                        */
```
$D' \longleftarrow \emptyset$
$D' \longleftarrow \mathrm{shuffle}(sample_{generic} \cup sample_{calibration})$
```
/* train the model h_m on dataset D'   */
```
$h_{m'} \longleftarrow h_m(D')$
return $h_{m'}$

We evaluated the classification models by computing their accuracy, precision, recall, and their F_1score when tested on the test datasets. As for the regression models, their performance is evaluated by calculating their mean absolute error (MAE) and their root mean squared error (RMSE).

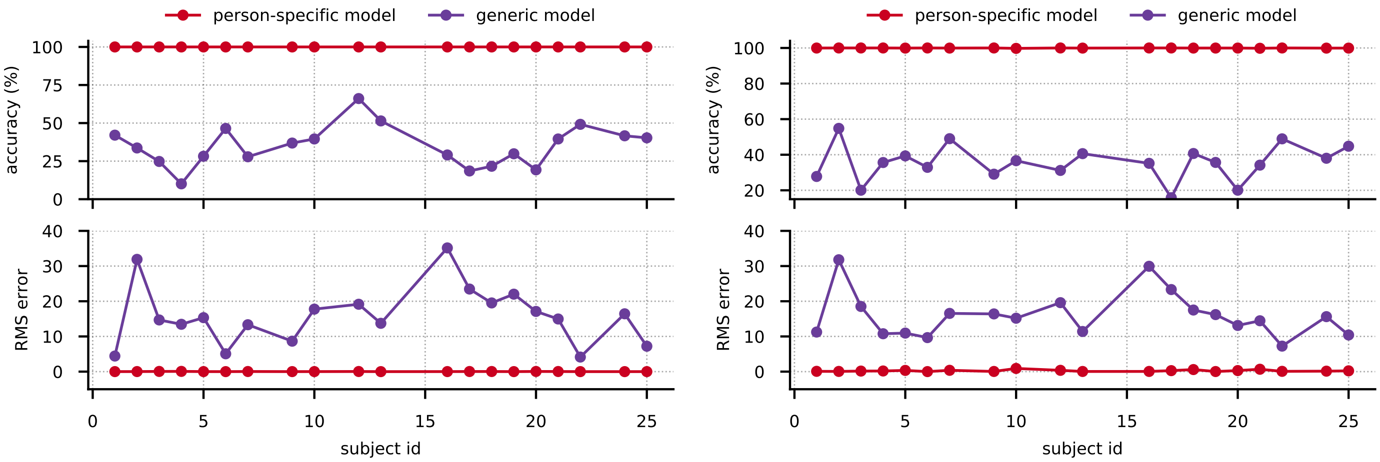
3 RESULTS AND DISCUSSION

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 unseen people. Indeed, the generic models (i.e., the models that could be used to predict the stress of any person) performed very poorly as shown in Figs. I and II.

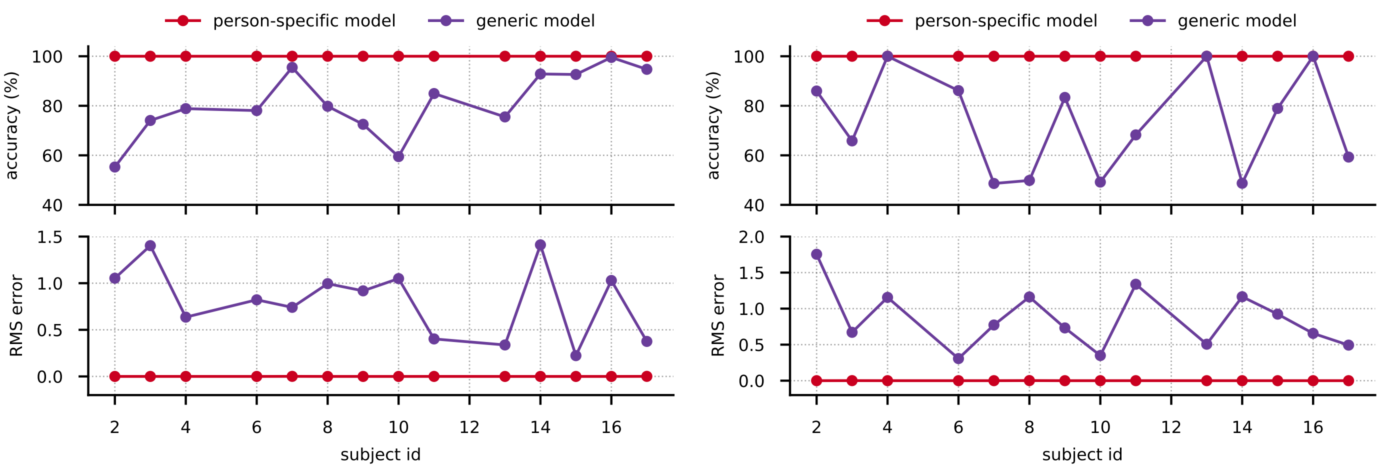
At first, we were dubious of this superb performance of the person-specific models, and we suspected the model were over-fitted. However, there is no indication that this was the case. First, we validated the models using a 10-fold cross-validation strategy, and it produced consistent predictions with a very low standard deviation between the 10-folds. Overfitting occurs when an ML model learns unnecessary details (e.g., noise) in the training set that are not available when making predictions on unseen data. K-Fold cross-validation provides an unbiased estimation of the performance of the model because it tests how well the k different parts of the training data perform on the model. Secondly, all the models use an extremely randomized trees model (Table III) that is less likely to overfit. Namely, the model is composed of a large number of shallow trees (500 trees, maximum depth=16) and a limited number of *best split features*. A low *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 a low generalization error variance and a high stability [63]. Moreover, the trees are shallow (maximum depth=16) to minimize overfitting. Thirdly, even simple person-specific models that underfit the training data performed better than the best generic model. For example, a person-specific decision stump(i.e., a one-level decision tree) model achieved an 80% accuracy. Finally, the discrepancies between the person-specific and generic models are nothing out of the ordinary because other prominent studies on stress prediction have achieved similar results. In general, person-specific models achieve stress prediction accuracy greater than 90% [47], [37] [40] [64] [18] [28] [36] while generic models always under-perform [37] [23] [36]. Our stress prediction method achieved slightly higher accuracy compared to most other published literature mainly because of the method we used to extract and process the features. For example, when extracting HRV features, our window-sliding HRV computation method (Section 2.2) allows us to track how each heartbeat influences a person's HRV meticulously and to pinpoint whether a given subject is stressed or not accurately.

The drop in accuracy, when tested on unseen subjects, is also nothing out of the ordinary because, as already suggested in Section 1, the predictive models cannot learn the inter-subject physiological differences in how people respond to the stressors. To test 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 a mean decrease in impurity (MDI)) of the *subject id* to 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; thus, is the most critical attribute for stress prediction. Additionally, as shown in



(a) SWELL HRV dataset **(b)** SWELL EDA dataset

FIG. I. Performance comparison between the person-specific and the generic models trained on the SWELL datasets For all subjects, the person-specific classification models consistently achieved a perfect or near-perfect accuracy, and the regression models have a near-zero RMS error. However, because of the inter-individual differences reacting to stress, all the generic models performed poorly, and there is a vast performance variation between the subjects.



(a) WESAD HRV dataset **(b)** WESAD EDA dataset

FIG. II. Performance comparison between the person-specific and the generic models trained on the WESAD datasets For all subjects, the person-specific classification models consistently achieved a perfect or near-perfect accuracy, and the regression models have a near-zero RMS error. However, because of the inter-individual differences reacting to stress, all the generic models performed poorly, and there is a vast performance variation between the subjects.

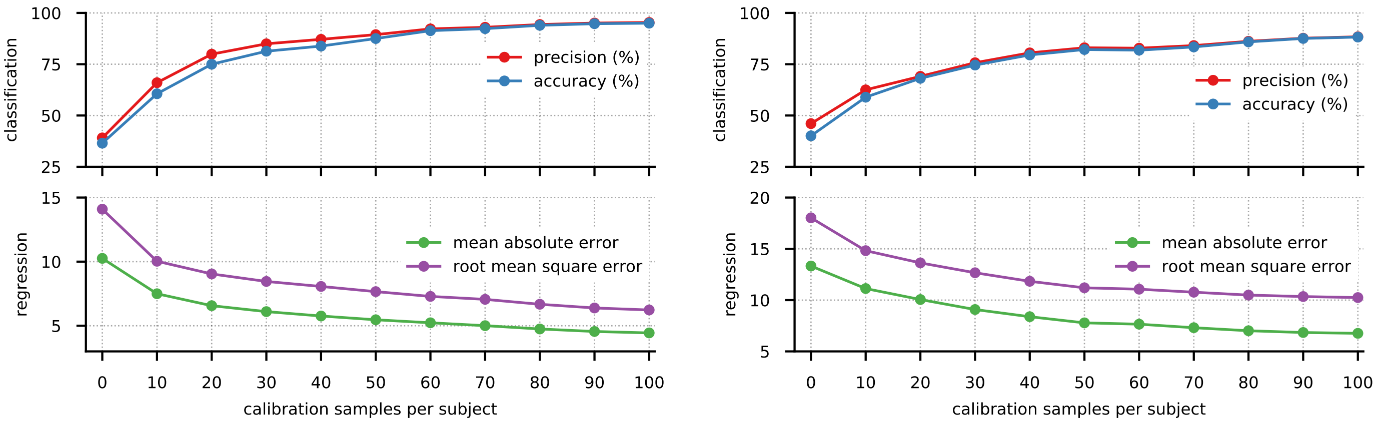
Figs. I and II, unlike the person-specific models, because each subject has a unique response to stress, the generic model's performance varies widely between the different subjects.Therefore, a one-size-fits-all stress prediction models will have an unpredictable and low performance compared to the 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 adequate hyperparameters optimizations), it was clear that the performance of the person-specific models dwarfs that of the person-independent models. Furthermore, the hyperparameters tuning is perverse guesswork and an erratic process because the distribution of each subject is somehow unique; therefore, finding compromising hyperparameters for a model that works for all subjects is a futile endeavor.

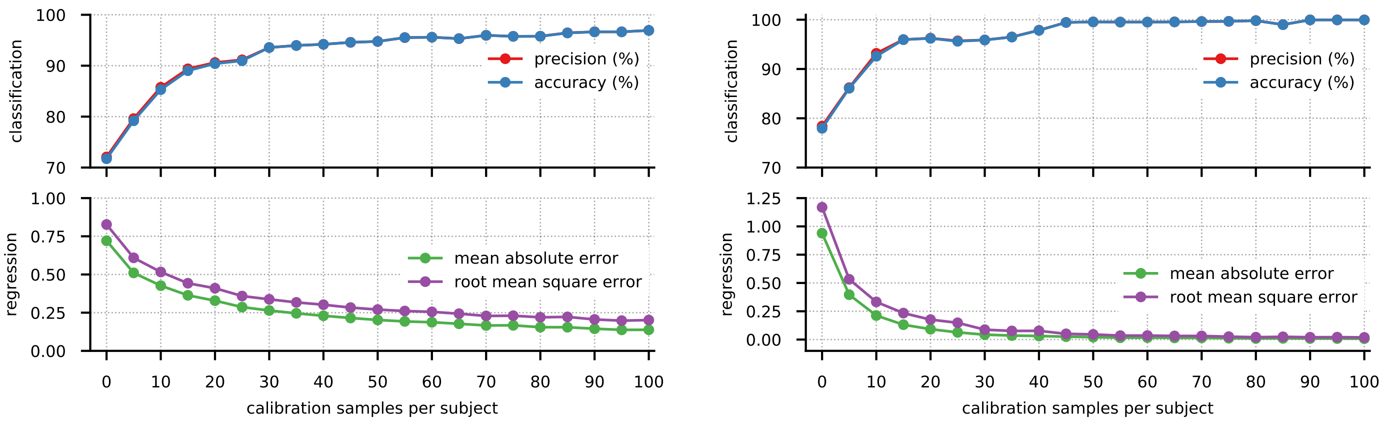
In an attempt to improve the models' generalization on unseen people, we investigated how each model would perform if it knew little information about the previously unseen subjects. Consequently, we devised a technique that derive a person-specific like model from the data collected from a large group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

When we applied this technique to stress prediction on the two datasets, the performance of all the models signifi-



(a) SWELL HRV dataset **(b)** SWELL EDA dataset

FIG. III. Performance of the hybrid model trained on the SWELL dataset
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 dataset **(b)** WESAD EDA dataset

FIG. IV. Performance of the hybrid model trained on the WESAD dataset
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

cantly increased, even when we only used a few calibration samples (Figs. III and IV).

- For the regression models, the root-mean-square error (RMSE) and mean absolute error (MAE) sharply dropped when we used a few calibration samples, and this is the case both for 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 mean absolute error decreased from 10.18 to 7.45 when only 10 calibration samples per unseen subject were used. Likewise, this error dropped even more so when 100 bootstrapping samples were added to the training set (mean absolute error =4.42, root-mean-square error=6.31).
- In a like manner, the performance of the classification models noticeably increased when a few bootstrapping seed samples were added to their training datasets. For instance, the model trained on the HRV signal of the SWELL dataset, the accuracy, the precision and recall respectively increased from 57.77%, 38.20%, and 36.00% to 73.93%, 67.68% and 60.19% when only 10 bootstrapping samples per unseen subjects are added to the training set and culminated in a 96.59% accuracy, 95.37% precision and 94.66% recall when 100 bootstrapping samples per subject were added.

The increase in performance due to the few person-specific calibration samples highlights the influence of the person-specific biometrics in predicting stress. In [48], the authors showed that, when inter-individual physiological differences are not accounted for, stress predictive models may perform no better than models with no learning capability. Our result highly their findings. Nevertheless, all humans share a common hormonal response to stress [53], albeit a person's unique factors such as gender [39], genetics [54], personality [43], weight [55] and his coping ability [38] differentiate how each person reacts to stress. Previous researchers (e.g., [23], [50], [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-product stress monitoring product because they rely on heuristic clustering methods, and there is no authoritative subject clustering criterion. Our proposed method is simpler and much cheaper for a real-world deployment (see discussion in Section 4) and performs much better than any previously proposed generic model improvement method.

4 STRESS MONITORING IN OFFICES

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

The architecture and deployment of a continuous stress monitoring system that uses our technique will undoubtedly involve a lot of technical challenges that are beyond the scope of this paper. The interested reader is encouraged to read [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 an adequate sampling frequency because the quality of the physiological data affects the performance of the stress prediction models [65]. These stringent requirements necessitate making conflicting compromises. For example, 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. There exist many wearable devices (e.g., smart-watches and fitness trackers) with built-in PPG sensors. For example, the Empatica E4 wristband[2] might serve for this purpose. The device boasts of a high-resolutionEDA sensor with a durable steel electrode that can continuously record both the tonic and phasic changes in the skin conductance. As discussed in a recent article [66], the Empatica E4 wrist band has a satisfactory accuracy in recording HRV in 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 [3], [5] 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. The simples, and likely, 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 consumed to predict the person's stress. Regrettably, such an approach is daunting, time-consuming, and expensive because it requires to collect, label, and clean the data of each person. For example, in an office environment, this necessitates collecting, labeling, and training new data for every employee. Moreover, once deployed, the resulting stress monitoring system will unquestionably not perform as expected because its performance would deteriorate with time because a person's stress is dynamic and affected by many factors [52], [67]. 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 the system from the anticipated performance degradation.

As suggested by the results of this paper, an alternative and cost-effective method would be to use a generic model (i.e., a model trained on a large population) as a baseline and further tweaked it with data recorded from each user. With this approach, it could be possible to use the physiological fingerprints (i.e., tinny physiological samples) of each person to create a personalized model derived from the data of a large group of people. As an illustration, after training and testing a generic stress prediction model, it could be possible to create an automatic self-updating stress prediction model pipeline (Figure V) as follows:

1) *Step 1: personalized data collection* —once the stress monitoring system is deployed for the first time (at this point, it uses only a generic model), it is primordial that its users take several self-evaluation surveys in different working conditions to allow the collection of self-evaluation ground-truths that reflect the broader rangers of stressors the users are exposed to. 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" samples from the users, it would automatically create each user's personalized model by training a new model on a combination of the new user-specific data with the data that was used to train the generic model.
2) *Step 2: continuous machine learning* —Once this personalized model is deployed, its users are reminded to collect addition "calibration" ground truth data by taking shorts self-report survey periodically (e.g., via a web survey every time he/she finished a task) 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 [68], i.e., they will become stale because their input data unpredictably change over time. In stress prediction, model drifting is particularly inevitable because stress is inherently dynamic [67]. The models, thus need to adapt to the new changes. For example, when the system has received a specific number of new "calibration" self-assessment samples from a user, it would automatically test their accuracy against the existing model. If this prediction indicates a deterioration of the model, the system will need to update the model to cancel the model 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 be, however, computationally expensive and require significant time to retrain each user's model. Depending on the system, it would be instead more appropriate to incrementally train the existing

2. https://www.empatica.com/research/e4/

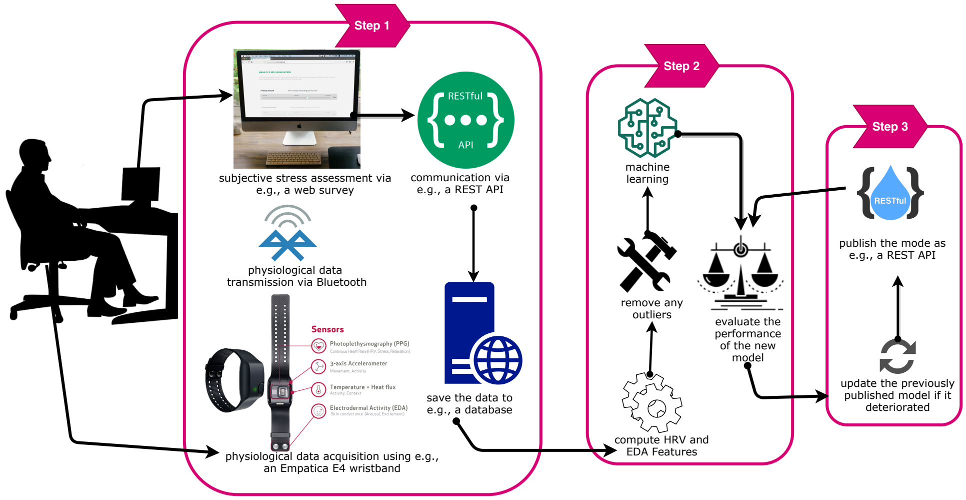


FIG. V. A simplified pipeline for a continuous stress monitoring model —a person's photoplethysmogram (PPG) and electrodermal activity (EDA) signals are recorded using a wristband device. The signals are sent to a computing device where appropriate features (e.g., Table I) are computed, preprocessed (e.g., data cleaning, rebalancing and dimension reduction) and sent to a remote server where they are used to predict the person's stress. For calibration purses, the person also periodically provide self-assessment of his stress (e.g., via a web survey after the completion of his work). This feedback is used to train a personalized stress prediction model which is published and consumed as a RESTful API. When the model deteriorates, it is automatically updated based on the periodic self-evaluations the system received from its users.

model as the new data is received [69]. This approach is faster because it does not require retraining the whole model when new data comes. Instead, it extends the existing model by, e.g., combining the new data with a subset of the old data [70]. Nevertheless, It is important to note that many machine learning algorithms do not support incremental learning and that, unless there is rigorous monitoring of the system, incremental learning may introduce nefarious predicaments [69].

3) *Step 3: publication of the model* —the model is published as, e.g., a REST Application Program Interface (REST API) and periodically updated depending on its performance.

Using this approach presents 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 require expensive installation, support, and maintenance services costs. Our approach would likely be less expensive because there will be no need to collect large quantities of new data from each user. Instead, only a few user-specific samples are 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 subject-dependent 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, would be very costly to implement, and, most importantly, this approach is rigid and not flexible to the expected dynamic changes in each user's dynamic stress changes. The proposed approach achieves a stress prediction accuracy that is comparable to that achieved by subject-dependent models and yet, presents enticing large scale deployment benefits.
- *straightforward deployment* —once deployed, each user's person-specific model can be generated using negligible user-specific samples that can be unobtrusively collected using, e.g., an approach proposed in [71] in which each user can self-evaluate (in terms of NASA-TLX and SSSQ) his stress level via a smartphone application. The self-evaluation would serve as a person-specific calibration to the generic model. Over time, when the model degrades due to the person's dynamics in stress, 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, there are still 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. Further, 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 it is often heralded as one of the best indicators of stress [5], [72], it has significant drawbacks. The EDA is a result of electrical changes that happen when the skin receives signals from the nervous system. Under stress, the skin's conductance changes due to a subtle increase in sweat that lead 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, the ambient temperature, and the ambient humidity. Moreover, for the same person, an EDA signal may fluctuate from one day to another [73]. Additionally, because stress is intrinsically multifaceted (it consists of physiological, behavioral and affective response), as highlighted in [74], it is imperative to take into consideration its context (i.e., where, what, when, who, why, and how). This approach, as shown in [75], may yield better and predictable results even when tested in 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 out of a budget of a small business. However, the investment might be well worth it for a large business. In our previous studies, we showed that it is possible to predict people's thermal comfort using the variations in their HRV [76] [77], and highlighted the energy saving potential of this approach [78]. Therefore, the positive spillovers that might result in using the system may outstrip the initial investment because, in a responsive smart office, the system can be used as part a of a multipurpose system that uses the office occupants' physiological signals for preventive medicine, stress management, and provides an efficient thermal comfort at low energy. Additionally, there exist enabling technologies that would make these challenges a little bit easier. For example, IBM's Watson Studio[3] offers tools that simplify developing and deploying predictive models. In our proposed stress monitoring system, Watson Studio could be used —and requires little or no programming experience —to automate steps 1 and 2 (see Figure V) including model deterioration monitoring and deployment as a REST API.

5 CONCLUSION

Despite an extensive body of literature on stress recognition, and notwithstanding the potential economic and health benefit of stress monitoring, there exists today no robust real-world stress recognition system. The most reliable and uncompromising methods use a fusion of multi-modal signals (e.g., physiological (such as HRV, EDA, EEG, EMG, skin temperature, respiration, pupil diameters, eye gaze), behavioral (keystrokes and mouse dynamics, and sitting posture), facial expression, speech patterns, and mobile phone use patterns). This approach, however, raises both practical challenges (e.g., real-time multi-modal data acquisition, data fusion, and data integration) and user privacy concerns (e.g., the implication of recording a person's computer keystrokes, his video and his speech), and, are not feasible in the real-world settings because of company-wide computer security policies or due to international workplace privacy laws.

On the contrary, the most practical stress monitoring methods use physiological signals. Physiological signals well in predicting stress, and unlike other methods, they are more reliable and harder to feign because the autonomous nervous system controls the fluctuation in physiological signals. Nevertheless, because stress is inherently subjective and is felt differently depending on the person, methods that use generic stress prediction models underperform when they are tested on 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 collecting new data and training a new model for every user of the system. In an office environment, this entails spending precious resources to collect and train a new model for every employee. Moreover, because stress is inherently dynamic, these models will need expensive periodic updates to collect and retrain every model to prevent the system from deterioration due to concept drift.

In this paper, we proposed to customize an individual's stress prediction model using a generic stress prediction model trained on a large population. Our method takes its foundation on the fact that humans share a common hormonal response to stress. However, every person possesses unique factors (e.g., gender, age, weight, and copying ability) that differentiate the person from others. Therefore, we hypothesized that it could be possible to improve the generalization performance of a generic stress prediction model trained on a large population by incorporating tinny biometric samples collected from previously unseen people into the generic model. We tested our method on two stress datasets and found that our approach significantly increased the generalization performance of the generic model. Furthermore, we surmised that it, in order to create a practical stress monitoring system, it could be possible to use small "fingerprint" samples collected from each user of the system and create a personalized model derived from the generic model. We argued that this approach would be cost-effective and practical to deploy in real-world settings. We also highlighted the technical limitations of the proposed method. Our future research will focus on tackling these issues.

SUPPLEMENTARY MATERIAL

Additional supporting information may be found online [4] in our public repository and contains more detailed information and source code to replicate our finding:

- Source code to reproduce the relevant findings of this research
- Dataset of the computed HRV and EDA features
- HRV and EDA feature importance with and without the *subject_id* added to the datasets (Section 2.3)
- Detailed tables of the performance of the person-specific models (Section 3.1)
- Detailed tables of the performance of the generic models (Section 3.1)
- Detailed tables of the of performance of the calibrated models (Section 3.2)

REFERENCES

[1] D. Carneiro, P. Novais, J. C. Augusto, and N. Payne, "New Methods for Stress Assessment and Monitoring at the Workplace," *IEEE Trans. Affect. Comput.*, vol. 10, no. 2, pp. 237–254, apr 2019.

[2] Y. S. Can, B. Arnrich, and C. Ersoy, "Stress detection in daily life scenarios using smart phones and wearable sensors: A survey," *J. Biomed. Inform.*, vol. 92, p. 103139, apr 2019.

[3] P. Schmidt, A. Reiss, R. Dürichen, and K. V. Laerhoven, "Wearable affect and stress recognition: A review," *CoRR*, vol. abs/1811.08854, 2018. [Online]. Available: http://arxiv.org/abs/1811.08854

[4] D. Carneiro, P. Novais, J. C. Augusto, and N. Payne, "New methods for stress assessment and monitoring at the workplace," *IEEE Trans. Affect. Comput.*, vol. 14, no. 8, pp. 1–1, 2017.

[5] A. Alberdi, A. Aztiria, and A. Basarab, "Towards an automatic early stress recognition system for office environments based on multimodal measurements: A review," *J. Biomed. Inform.*, vol. 59, pp. 49–75, 2016.

[6] E. D. Kirby, S. E. Muroy, W. G. Sun, D. Covarrubias, M. J. Leong, L. A. Barchas, and D. Kaufer, "Acute stress enhances adult rat hippocampal neurogenesis and activation of newborn neurons via secreted astrocytic fgf2," *Elife*, vol. 2013, no. 2, pp. 1–23, 2013.

[7] F. S. Dhabhar, W. B. Malarkey, E. Neri, and B. S. McEwen, "Stress-induced redistribution of immune cells-from barracks to boulevards to battlefields: A tale of three hormones - curt richter award winner," *Psychoneuroendocrinology*, vol. 37, no. 9, pp. 1345–1368, sep 2012.

[8] C. Tennant, "Work-related stress and depressive disorders," *J. Psychosom. Res.*, vol. 51, no. 5, pp. 697–704, 2001.

3. https://www.ibm.com/cloud/machine-learning

4. available at https://www.kaggle.com/qiriro/tac

[9] T. W. Colligan and E. M. Higgins, "Workplace stress: Etiology and consequences," *J. Workplace Behav. Health*, vol. 21, no. 2, pp. 89–97, 2005.

[10] D. C. Ganster and C. C. Rosen, "Work stress and employee health," *J. Manage.*, vol. 39, no. 5, pp. 1085–1122, jul 2013.

[11] EU-OSHA, "Psychosocial risks and stress at work - safety and health at work," 2017.

[12] J. M. Peake, G. Kerr, and J. P. Sullivan, "A critical review of consumer wearables, mobile applications, and equipment for providing biofeedback, monitoring stress, and sleep in physically active populations," *Front. Physiol.*, vol. 9, no. JUN, pp. 1–19, 2018.

[13] A. H. Marques, M. N. Silverman, and E. M. Sternberg, "Evaluation of stress systems by applying noninvasive methodologies: Measurements of neuroimmune biomarkers in the sweat, heart rate variability and salivary cortisol," *Neuroimmunomodulation*, vol. 17, no. 3, pp. 205–208, 2010.

[14] H. Hellhammer and C. Kirschbaum, "Salivary cortisol in psychoneuroendocrine research: recent developments and applications," *Psychoneuroendocrinology*, vol. 19, no. 4, pp. 313–333, 1994.

[15] K. P. Eisen, G. J. Allen, M. Bollash, and L. S. Pescatello, "Stress management in the workplace: A comparison of a computer-based and an in-person stress-management intervention," *Comput. Human Behav.*, vol. 24, no. 2, pp. 486–496, 2008.

[16] S. Järvelin-Pasanen, S. Sinikallio, and M. P. Tarvainen, "Heart rate variability and occupational stress-systematic review." *Ind. Health*, vol. 56, no. 6, pp. 500–511, nov 2018.

[17] P. Adams, M. Rabbi, T. Rahman, M. Matthews, A. Voida, G. Gay, T. Choudhury, and S. Voida, "Towards personal stress informatics: Comparing minimally invasive techniques for measuring daily stress in the wild," *Proc. 8th Int. Conf. Pervasive Comput. Technol. Healthc.*, pp. 72–79, 2014.

[18] P. Melillo, M. Bracale, and L. Pecchia, "Nonlinear heart rate variability features for real-life stress detection. case study: students under stress due to university examination," *Biomed. Eng. Online*, vol. 10, no. 1, p. 96, 2011.

[19] B. Cinaz, B. Arnrich, R. La Marca, and G. Tröster, "Monitoring of mental workload levels during an everyday life office-work scenario," *Pers. Ubiquitous Comput.*, vol. 17, no. 2, pp. 229–239, feb 2013.

[20] K. S. Rahnuma, A. Wahab, N. Kamaruddin, and H. Majid, "Eeg analysis for understanding stress based on affective model basis function," *Proc. Int. Symp. Consum. Electron. ISCE*, pp. 592–597, 2011.

[21] C. Z. Wei, "Stress emotion recognition based on rsp and emg signals," *Adv. Mater. Res.*, vol. 709, pp. 827–831, jun 2013.

[22] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Inf. Fusion*, vol. 37, pp. 98–125, 2017.

[23] S. Koldijk, M. A. Neerincx, and W. Kraaij, "Detecting work stress in offices by combining unobtrusive sensors," *IEEE Trans. Affect. Comput.*, vol. 9, no. 2, pp. 227–239, 2018.

[24] O. M. Mozos, V. Sandulescu, S. Andrews, D. Ellis, N. Bellotto, R. Dobrescu, and J. M. Ferrandez, "Stress detection using wearable physiological and sociometric sensors," *Int. J. Neural Syst.*, vol. 27, no. 02, p. 1650041, 2017.

[25] M. Gjoreski, H. Gjoreski, M. Luštrek, and M. Gams, "Continuous stress detection using a wrist device," in *Proc. 2016 ACM Int. Jt. Conf. Pervasive Ubiquitous Comput. Adjun. - UbiComp '16*. New York, New York, USA: ACM Press, 2016, pp. 1185–1193.

[26] R. Kocielnik, N. Sidorova, F. M. Maggi, M. Ouwerkerk, and J. H. D. M. Westerink, "Smart technologies for long-term stress monitoring at work," *Proc. CBMS 2013 - 26th IEEE Int. Symp. Comput. Med. Syst.*, 2013.

[27] J. Zhai and A. Barreto, "Stress detection in computer users through non-invasive monitoring of physiological signals," *Biomed. Sci. Instrum.*, vol. 42, pp. 495–500, 2006.

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

[29] R. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: analysis of affective physiological state," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 10, pp. 1175–1191, 2001.

[30] A. Haag, S. Goronzy, P. Schaich, and J. Williams, "Emotion recognition using bio-sensors: First steps towards an automatic system," in *Tutor. Res. Work. Affect. dialogue Syst.*, 2004, vol. i, pp. 36–48.

[31] G. Valenza, L. Citi, A. Lanatá, E. P. Scilingo, and R. Barbieri, "Revealing real-time emotional responses: A personalized assessment based on heartbeat dynamics," *Sci. Rep.*, vol. 4, pp. 1–13, 2014.

[32] A. Alberdi, A. Aztiria, A. Basarab, and D. J. Cook, "Using smart offices to predict occupational stress," *Int. J. Ind. Ergon.*, vol. 67, no. April, pp. 13–26, 2018.

[33] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing wesad, a multimodal dataset for wearable stress and affect detection," *Proc. 2018 Int. Conf. Multimodal Interact. - ICMI '18*, pp. 400–408, 2018.

[34] A. Zenonos, A. Khan, G. Kalogridis, S. Vatsikas, T. Lewis, and M. Sooriyabandara, "Healthyoffice: Mood recognition at work using smartphones and wearable sensors," in *2016 IEEE Int. Conf. Pervasive Comput. Commun. Work. (PerCom Work.* IEEE, mar 2016, pp. 1–6.

[35] V. Kolodyazhniy, S. D. Kreibig, J. J. Gross, W. T. Roth, and F. H. Wilhelm, "An affective computing approach to physiological emotion specificity: Toward subject-independent and stimulus-independent classification of film-induced emotions," *Psychophysiology*, vol. 48, no. 7, pp. 908–922, 2011.

[36] J. Kim and E. Andre, "Emotion recognition based on physiological changes in music listening," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 12, pp. 2067–2083, dec 2008.

[37] Y. Nakashima, J. Kim, S. Flutura, A. Seiderer, and E. André, "Stress Recognition in Daily Work," in *Pervasive Comput. Paradig. Ment. Heal.*, 2016, vol. 604, pp. 23–33.

[38] L. Kogler, V. I. Müller, A. Chang, S. B. Eickhoff, P. T. Fox, R. C. Gur, and B. Derntl, "Psychosocial versus physiological stress — meta-analyses on deactivations and activations of the neural correlates of stress reactions," *Neuroimage*, vol. 119, pp. 235–251, oct 2015.

[39] J. Wang, M. Korczykowski, H. Rao, Y. Fan, J. Pluta, R. C. Gur, B. S. McEwen, and J. A. Detre, "Gender difference in neural response to psychological stress," *Soc. Cogn. Affect. Neurosci.*, vol. 2, no. 3, pp. 227–239, sep 2007.

[40] A. Liapis, C. Katsanos, D. Sotiropoulos, M. Xenos, and N. Karousos, "Stress recognition in human-computer interaction using physiological and self-reported data: A study of gender differences," *ACM Int. Conf. Proceeding Ser.*, vol. 01-03-Octo, pp. 323–328, 2015.

[41] M. Matud, "Gender differences in stress and coping styles," *Pers. Individ. Dif.*, vol. 37, no. 7, pp. 1401–1415, nov 2004.

[42] J. Hernandez, R. R. Morris, and R. W. Picard, "Call center stress recognition with person-specific models," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6974 LNCS, no. PART 1, pp. 125–134, 2011.

[43] E. Childs, T. L. White, and H. de Wit, "Personality traits modulate emotional and physiological responses to stress," *Behav. Pharmacol.*, vol. 25, no. 5-6, p. 1, jul 2014.

[44] M. Johnstone and J. A. Feeney, "Individual differences in responses to workplace stress: The contribution of attachment theory," *J. Appl. Soc. Psychol.*, vol. 45, no. 7, pp. 412–424, 2015.

[45] R. M. Sapolsky, "Individual differences and the stress response," pp. 261–269, 1994.

[46] D. M. Almeida, J. R. Piazza, and R. S. Stawski, "Interindividual differences and intraindividual variability in the cortisol awakening response: An examination of age and gender." *Psychol. Aging*, vol. 24, no. 4, pp. 819–827, 2009.

[47] N. Attaran, A. Puranik, J. Brooks, and T. Mohsenin, "Embedded low-power processor for personalized stress detection," *IEEE Trans. Circuits Syst. II Express Briefs*, vol. 65, no. 12, pp. 2032–2036, 2018.

[48] B. Lamichhane, U. Großekathöfer, G. Schiavone, and P. Casale, "Towards Stress Detection in Real-Life Scenarios Using Wearable Sensors: Normalization Factor to Reduce Variability in Stress Physiology," 2017, vol. 181, pp. 259–270.

[49] J. Aigrain, "Multimodal detection of stress : evaluation of the impact of several assessment strategies," PhD Thesis, Universit´e Pierre et Marie Curie Ecole´, 2016.

[50] Q. Xu, T. L. Nwe, and C. Guan, "Cluster-Based Analysis for Personalized Stress," *Ieee J. Biomed. Heal. Informatics*, vol. 19, no. 1, pp. 275–281, 2015.

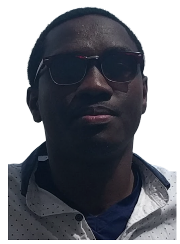
[51] J. Ramos, J.-h. Hong, and A. K. Dey, "Stress Recognition - A Step Outside the Lab," *Proc. 1st Int. Conf. Physiol. Comput. Syst. PhyCS 2014*, pp. 107–118, 2014.

[52] N. Schneiderman, G. Ironson, and S. D. Siegel, "Stress and health: Pyshological, behavior, and biological," *October*, vol. 1, no. Lacey 1967, pp. 1–19, 2008.

[53] E. Charmandari, C. Tsigos, and G. Chrousos, "Endocrinology of the stress response." *Annu. Rev. Physiol.*, vol. 67, no. 1, pp. 259–84, mar 2005.

[54] S. Wüst, I. S. Federenko, E. F. C. van Rossum, J. W. Koper, R. Kumsta, S. Entringer, and D. H. Hellhammer, "A psychobiological perspective on genetic determinants of hypothalamus-pituitary-

adrenal axis activity." *Ann. N. Y. Acad. Sci.*, vol. 1032, no. 1, pp. 52–62, dec 2004.
[55] S. U. Jayasinghe, S. J. Torres, C. A. Nowson, A. J. Tilbrook, and A. I. Turner, "Physiological responses to psychological stress: importance of adiposity in men aged 50–70 years," *Endocr. Connect.*, vol. 3, no. 3, pp. 110–119, 2014.
[56] M. Malik, J. T. Bigger, A. J. Camm, R. E. Kleiger, A. Malliani, A. J. Moss, and P. J. Schwartz, "Heart rate variability: Standards of measurement, physiological interpretation, and clinical use," *Eur. Heart J.*, vol. 17, no. 3, pp. 354–381, mar 1996.
[57] R. Zangróniz, A. Martinez-Rodrigo, J. M. Pastor, M. T. López, and A. Fernández-Caballero, "Electrodermal activity sensor for classification of calm/distress condition," *Sensors*, vol. 17, no. 10, pp. 1–14, 2017.
[58] S. Koldijk, M. Sappelli, S. Verberne, M. A. Neerincx, and W. Kraaij, "The swell knowledge work dataset for stress and user modeling research," *Proc. 16th Int. Conf. Multimodal Interact. - ICMI '14*, pp. 291–298, 2014.
[59] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," *Adv. Psychol.*, vol. 52, no. C, pp. 139–183, jan 1988.
[60] C. Kirschbaum, K. M. Pirke, and D. H. Hellhammer, "The 'trier social stress test'–a tool for investigating psychobiological stress responses in a laboratory setting." *Neuropsychobiology*, vol. 28, no. 1-2, pp. 76–81, jun 1993.
[61] W. S. Helton and K. Näswall, "Short stress state questionnaire: Factor structure and state change assessment." *Eur. J. Psychol. Assess.*, vol. 31, no. 1, pp. 20–30, 2015.
[62] I.-K. Yeo, "A new family of power transformations to improve normality or symmetry," *Biometrika*, vol. 87, no. 4, pp. 954–959, dec 2000.
[63] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
[64] G. Rigas, Y. Goletsis, and D. I. Fotiadis, "Real-time driver's stress event detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 1, pp. 221–234, 2012.
[65] A. Chowdhury, R. Shankaran, M. Kavakli, and M. M. Haque, "Sensor Applications and Physiological Features in Drivers' Drowsiness Detection: A Review," *IEEE Sens. J.*, vol. 18, no. 8, pp. 3055–3067, 2018.
[66] L. Menghini, E. Gianfranchi, N. Cellini, E. Patron, M. Tagliabue, and M. Sarlo, "Stressing the accuracy: Wrist-worn wearable sensor validation over different conditions," *Psychophysiology*, no. December 2018, pp. 1–15, jul 2019.
[67] E. O. Johnson, T. C. Kamilaris, G. P. Chrousos, and P. W. Gold, "Mechanisms of stress: a dynamic overview of hormonal and behavioral homeostasis." *Neurosci. Biobehav. Rev.*, vol. 16, no. 2, pp. 115–30, 1992.
[68] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM Comput. Surv.*, vol. 46, no. 4, pp. 1–37, mar 2014.
[69] A. Gepperth and B. Hammer, "Incremental learning algorithms and applications," *Eur. Symp. Artif. Neural Networks*, no. April, pp. 357–368, 2016.
[70] F. M. Castro, M. J. Marín-Jiménez, N. Guil, C. Schmid, and K. Alahari, "End-to-end incremental learning," in *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, 2018, vol. 11216 LNCS, pp. 241–257.
[71] N. E. Bush, G. Ouellette, and J. Kinn, "Utility of the t2 mood tracker mobile application among army warrior transition unit service members," *Mil. Med.*, vol. 179, no. 12, pp. 1453–1457, 2014.
[72] C. Setz, B. Arnrich, J. Schumm, R. La Marca, G. Troster, and U. Ehlert, "Discriminating stress from cognitive load using a wearable eda device," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 2, pp. 410–417, mar 2010.
[76] K. Nkurikiyeyezu and G. Lopez, "Toward a real-time and physiologically controlled thermal comfort provision in office buildings," in *Intell. Environ.* IOS Press, 2018, pp. 168–177.
[73] J. Bakker, M. Pechenizkiy, and N. Sidorova, "What's your current stress level? detection of stress patterns from gsr sensor data," *Proc. - IEEE Int. Conf. Data Mining, ICDM*, no. 1, pp. 573–580, 2011.
[74] Y. Panagakis, O. Rudovic, and M. Pantic, "Learning for multimodal and affect-sensitive interfaces," in *Handb. Multimodal-Multisensor Interfaces Found. User Model. Common Modality Comb. - Vol. 2.* Association for Computing Machinery, oct 2018, pp. 71–98.
[75] M. Gjoreski, M. Luštrek, M. Gams, and H. Gjoreski, "Monitoring stress with a wrist device using context," *J. Biomed. Inform.*, vol. 73, pp. 159–170, 2017.
[77] K. N. Nkurikiyeyezu, Y. Suzuki, and G. F. Lopez, "Heart rate variability as a predictive biomarker of thermal comfort," *J. Ambient Intell. Humaniz. Comput.*, vol. 9, no. 5, pp. 1465–1477, aug 2018.
[78] K. Nkurikiyeyezu, Y. Suzuki, P. Maret, G. Lopez, and K. Itao, "Conceptual design of a collective energy-efficient physiologically-controlled system for thermal comfort delivery in an office environment," *SICE J. Control. Meas. Syst. Integr.*, vol. 11, no. 4, pp. 312–320, 2018.



Kizito Nkurikiyeyezu is a Ph.D. candidate at Aoyama Gakuin University, Tokyo, Japan. His doctoral research takes a multidisciplinary approach and investigates the possibility to estimate a person's thermal comfort level from the fluctuations in his physiological signals and to use appropriate constrained optimization algorithms to provide an optimum and personalized thermal comfort using the least possible energy. He received his B.Sc. in Electrical Engineering and an M.Sc. in Electrical & Computer Engineering degrees from Oklahoma Christian University, Okla. City, U.S. He is a member of the Information Processing Society of Japan (IPSJ), the Society of Instrument and Control Engineers (SICE) and Institute of Electrical and Electronics Engineers (IEEE)

Anna Yokokubo She received the M.S. (2012) in HCI Human-Computer Interaction (HCI) 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 Univ., Kanagawa Japan.

Guillaume Lopez He received an M.E. in Computer Engineering from INSA Lyon, France, an M.Sc. and a Ph.D. in Environmental Studies from the University of Tokyo, Japan in 2000, 2002, and 2005 respectively. He worked as a research engineer at Nissan Motor Corp. from September 2005, 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 AHI, IEEE, IPSJ, JSME.