Personalized Stress Detection with Self-supervised Learned Features

Stefan Matthes [*1] **Zhiwei Han** [*1] **Tianming Qiu** [1] **Bruno Michel** [2] **Sören Klinger** [1] **Hao Shen** [1] **Yuanting Liu** [1] **Bashar Altakrouri** [3]

Abstract

Automated stress detection using physiological sensors is challenging due to inaccurate labeling and individual bias in the sensor data. Previous methods consider stress detection as a supervised classification task, where bad labeling leads to a large performance drop. Furthermore, the poor generalizability to unseen subjects reveals the importance of personalizing stress detection for both inter- and intra-individual sensor data variability. Towards this end we present a label-free feature extractor and an efficient personalization method with the "human in the loop" approach. First, we capture the intra-individual variability and encode it in self-supervised learned features, which are usually well separable and independent of noisy stress labels. Next, personalization is achieved by assigning labels to critical reference points via very few interactions between subject and wearable device. The promising results of the conducted experiments show the effectiveness and efficiency of our proposed method.

1. Introduction

Stress is the difficulty of organisms to maintain homeostasis under stimuli that causes an imbalance in the autonomic nervous system (Horowitz, 1976). Stressors are categorized into two main classes, namely absolute and relative stressors (Lupien et al., 2007). Absolute stressors refer to events or situations that pose a real threat to life or well-being. In contrast, relative stressors refer to psychological threats that are induced by the individual's subjective interpretation of the situation in terms of its unpredictability, uncontrollability (Mason, 1968), or "social evaluative threat" (Dickerson & Kemeny, 2004). Psychological stress is induced by time pressure, excessive interruptions (Koldijk et al., 2014) or the imbalance between external demands and the individual's resources (Hobfoll, 1989). Stress up to a specific level can be considered positive so as to enhance alertness and productivity, while high stress usually leads to reduced productivity (Benson & Allen, 1980), impaired decision-making capabilities (Baddeley, 1972), decreased situational awareness (Vidulich et al., 1994), as well as life-threatening symptoms, e.g. improper decisions in fire fighting. Thus, the precise detection of unbearable stress contributes to a reliable stress management mechanism, improved team performance and reduced risk for individuals in dangerous missions.

Stress causes physiological responses which are reflected in many biosignals. In previous work, heart rate variability (HRV) has been proven a stable indicator for stress detection (Rodrigues et al., 2018; Huysmans et al., 2018). Other biosignals such as skin conductance (EDA) and muscle tension (EMG) (Oskooei et al., 2019) also provide a reliable measure over stress levels. Due to the importance of stress in our daily lives and the rapid recent development of wearable sensory systems (Schmidt et al., 2018b), a number of research projects (Schmidt et al., 2018a; Zenonos et al., 2016; Kolodyazhniy et al., 2011; Healey & Picard, 2005; Bogomolov et al., 2014; Picard et al., 2001; Valenza et al., 2014) have been carried out to enable automatic stress detection by algorithms.

Although person-specific models (Healey & Picard, 2005; Bogomolov et al., 2014; Picard et al., 2001; Valenza et al., 2014) can show superior performance over generic methods (Schmidt et al., 2018a; Zenonos et al., 2016; Kolodyazhniy et al., 2011) on selected subjects, they still fail in capturing generalized inter-individual differences. As a consequence, person-specific models do not generalize well in predicting stress in yet unseen subjects. The adaptation to new subjects usually requires complete retraining of the machine learning model, which is time-consuming and impractical in real-world applications. How to perform efficient personalization

[*]Equal contribution [1]fortiss GmbH [2]IBM Zurich Research Lab, Zurich, Switzerland [3]IBM Deutschland GmbH, Munich, Germany. Correspondence to: Stefan Matthes <matthes@fortiss.org>, Zhiwei Han <han@fortiss.org>, Tianming Qiu <qiu@fortiss.org>, Bruno Michel <bmi@zurich.ibm.com>, Sören Klinger <klinger@fortiss.org>, Hao Shen <shen@fortiss.org>, Yuanting Liu <liu@fortiss.org>, Bashar Altakrouri <altakrouri@de.ibm.com>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 108, 2020.

while maintaining the intra-individual difference remains a key challenge for automated stress detection.

In this work, we substantiate the need to consider individual differences between users and propose an efficient personalization approach for stress detection. Our approach utilizes self-supervised learned (SSL) features, which help to avoid the bias induced by unreliable labels in datasets. Subsequently, personalization is achieved by recalibration with efficient human-computer interaction. To show the effectiveness of the proposed method, we conducted experiments and validated the final results following the leave-one-subject-out cross-validation (LOSO-CV). For the sake of simplicity, we focus in this paper on heartbeat-related features. Our main contributions are the following:

- We show that taking more samples from training subjects can be harmful for capturing inter-individual variability between subjects, if there is limited access to data of the test subject.
- We introduce SSL features for personalization, which significantly reduce the number of required interactions.
- We propose an efficient personalization method and demonstrate its competitiveness experimentally by comparing the results with baselines.

2. Related Work

Modern methods for automated stress detection use a combination of different data sources such as video, audio, and physiological data (Poria et al., 2017; Koldijk et al., 2016; Mozos et al., 2017). Due to the importance of stress recognition in daily life and the increasing popularity of wearable devices, a number of research projects developed automated stress detection approaches based on biosignals collected by wearable devices (Carneiro et al., 2017; Can et al., 2019; Gjoreski et al., 2016). Several well-studied physiological signals provide reliable indicators of stress. The most prevalent include electrocardiography (HRV, heart activity) (Melillo et al., 2011; Munla et al., 2015; Boonnithi & Phongsuphap, 2011), electroencephalogram (EEG, brain activity) (Costin et al., 2012; Vanitha & Krishnan, 2016), electrodermal activity (EDA, skin conductance) (Setz et al., 2009; Ayzenberg et al., 2012), photoplethysmogram (PPG, blood activity) (Mokhayeri et al., 2011) and electromyography (EMG, muscle activity) (Wijsman et al., 2010).

Most machine learning based stress detection methods (Ghaderi et al., 2015; Zhai & Barreto, 2006) use a supervised classification approach. However, the poor label quality caused by non-standardized definitions of stress and varying stress resilience degrade detection performance. Although generic models (Ghaderi et al., 2015; Zhai & Barreto, 2006) showed good results in detecting stress when evaluated with a k-fold cross-validation on mixed data from all subjects, this setup is still unrealistic in real world applications because generic models suffer from severe generalization problems on unseen subjects (Nkurikiyeyezu et al., 2019; Koldijk et al., 2016). The main reason is that the data from the training set and the test subjects are not independent and identically distributed. This lack of i.i.d. strongly limits the generic models' generalization power (Xu & Mannor, 2012). As a result, a practical stress detection model should be able to identify inter- as well as intra-individual differences induced by different physical and psychological conditions such as gender, age, individual stress tolerance, and health status, which influence how humans experience stress (Nkurikiyeyezu et al., 2019).

An efficient way to achieve better generalizability is to personalize stress detection by capturing inter-individual differences. Stress detection personalization methods can be categorized into three classes. (Sharma & Gedeon, 2011; Sztyler & Stuckenschmidt, 2017) divided the participants into different user groups according to their profiles and trained a personalized model for each group. But the huge data amount required for the identification of correlations between user-profiles and personalization models prevents this method to be widely applied in real-world applications. The combination of user modeling and stress detection personalization is thus still an open topic.

(Shi et al., 2010) normalized the measurements of each participant and created a personalized version of a support vector machine. The normalization was done by subtracting the mean value of the recorded measurements while the participants were in a stress-free state. This approach requires little interaction at the beginning of initialization. We refer to this approach in the following as baseline calibration.

(Reiss & Stricker, 2013; Nkurikiyeyezu et al., 2019) created a mixed dataset by adding a small number of additional person-specific calibration points to a generic training set. While incorporating additional calibration samples from the test subject obviously leads to an improvement in classification performance, our experiments show that this method overemphasizes the training subjects. We find that the inclusion of additional samples from the $\mathbf{N}-1$ training subjects could actually lead to a reduction in performance if at least one person-specific measurement per condition is available.

In contrast to previous work, we present an adaptive approach to stress detection that allows for corrections through user feedback. In order to reduce the number of interactions for a better user experience, low-dimensional but representative physiological features are required for efficient personalization. Therefore, we introduce self-supervised learned (SSL) features derived from RR-intervals. Self-supervised learning (Jing & Tian, 2020) is a representation learning

method, which learns representation by leveraging labels created by self-supervision for free. SSL is widely applied in several domains e.g., computer vision (Gidaris et al., 2018; Zhang et al., 2016; Doersch et al., 2015), reinforcement learning (Srinivas et al., 2020) and audio recognition (Oord et al., 2018). There are two main advantages to using SSL features in the proposed method:

- More representative features are obtained by exploiting the intrinsic structure of the raw data.
- The generated low-dimensional features require fewer data points for personalization and can be better visualized compared to HRV parameters (typically about 15 dimensions).

In the next sections, we will compare the aforementioned methods on two publicly available datasets and our own dataset and demonstrate the efficiency of our method.

3. Approach

In this section, we introduce our SSL feature-based personalization approach for stress detection. The proposed model learns the underlying intra-individual differences in the training subjects and has the advantage of fast adaptation to unseen subjects in real-world applications. In the first subsection, we discuss the feature extraction procedure that is used to derive the SSL features as well as the HRV parameters. We then present an efficient personalization method for the stress detection model that aims at fast adaptation to new subjects by capturing the inter-individual variability in an interactive fashion.

3.1. Feature Extraction

As mentioned before, we focus on the electrocardiogram signal and its associated features. Nevertheless, the proposed methods can also be applied to other biosignals. After calculating the R-R intervals (time intervals between successive heartbeats) and removing outliers, we extracted HRV and SSL features based on segments consisting of 120 R-R intervals. A summary of the selected HRV features is shown in Table 1.

Besides HRV parameters, we extracted features with a self-supervised model. The model was trained with a contrastive loss which minimizes the distance between anchor points and positive examples while maximizing the distance between anchor points and negative examples in feature space. For a randomly chosen anchor segment x_a, we selected the corresponding positive segment x_p by randomly shifting the anchor segment by up to ± 30 R-R intervals and inverting it with a probability of 0.5. The negative segments x_{n_i} were randomly picked outside this region. Per anchor segment, we used 63 negative segments. The loss is given by

$$L = -\log\left(\frac{\exp(-\|f_a - f_p\|_2^2)}{\exp(-\|f_a - f_p\|_2^2) + \sum_i \exp(-\|f_a - f_{n_i}\|_2^2)}\right), \tag{1}$$

where f_a, f_p and $f_{n_i} \in \mathbb{R}^{\mathrm{d}}$ are the features of the anchor, positive and negative segments, respectively.

We computed the features with a convolutional neural network consisting of three convolutional blocks, followed by a fully connected layer. In Figure 1, we visualize the learned features for three randomly selected subjects and the same features after merging the individual data for $\mathrm{d} = 3$. In the following experiments we set the dimension of the embedding space to $\mathrm{d} = 5$.

Each convolutional block consists of a 10-filter convolution with kernel size 3, a ReLU activation function, and a max-pooling layer with size 2. The model was trained for $5 \cdot 10^5$ iterations using stochastic gradient descent (SGD) with a batch size of 64 and a learning rate of 10^{-4}.

Table 1. Selected heart rate variability features

Feature	Description
$\mathrm{RR}_{\mathrm{mean}}$	Mean of the R-R intervals
$\mathrm{RR}_{\mathrm{std}}$	Standard deviation of the R-R intervals
$\mathrm{RR}_{\mathrm{min}}$	Minimum R-R interval
$\mathrm{RR}_{\mathrm{max}}$	Maximum R-R interval
$\mathrm{RR}_{\mathrm{med}}$	Median R-R interval
$\mathrm{RR}_{\mathrm{rng}}$	Range of the R-R intervals
pNN20	% of intervals differing more than 20 ms
pNN50	% of intervals differing more than 50 ms
RMSSD	Root mean square of differences between successive R-R intervals
SDSD	Standard deviation of differences between successive R-R intervals
SD1	Short-term Poincare plot descriptor
SD2	Long-term Poincare plot descriptor
LF	Absolute power of the low-frequency band
HF	Absolute power of the high-frequency band
LF/HF	Ratio of LF-to-HF power
$\mathrm{LF}_{\mathrm{nu}}$	Normalized LF component
$\mathrm{HF}_{\mathrm{nu}}$	Normalized HF component

3.2. Personalization via Human-in-the-Loop

The individual physiological baselines of humans as well as the physiological responses to emotion-related stimuli are highly biased by inter-individual variability (Giakoumis et al., 2013). However, the previous person-specific research efforts suffer from limited efficiency compared to the person-independent solutions (Nkurikiyeyezu et al., 2019).

To tackle the inter-individual variability, we personalized stress detection via an efficient interactive process with the human in the loop. To the best of our knowledge, this is the first attempt to combine interactive machine learning

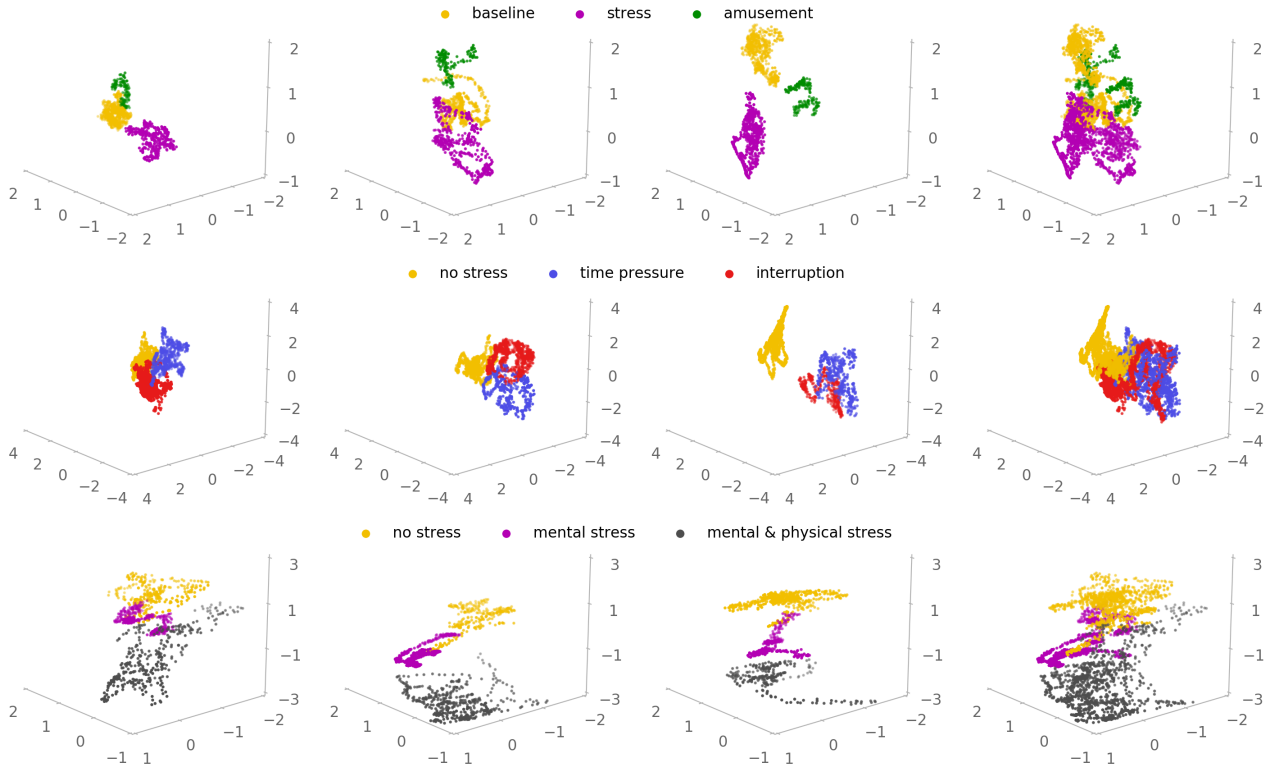

Figure 1. Self-supervised learned features for 3 random subjects (first 3 columns) and merged features (right column) for the WESAD (top row), SWELL (center row) and Eiken dataset (bottom row). Note that the feature dimension is set to 3 only for visualization purposes and might hide distinctive features due to loss of information

(Fails & Olsen Jr, 2003) with stress detection for better generalization of stress detection models. Our algorithm maintains a list of reference points, which we initialize with the feature means of each class and train our base model on those K (K is the number of unique labels) reference points in a supervised way. To capture the inter-individual difference of the current user, a few calibration points are collected from human-computer-interactions using different strategies. The user can either manually label the current stress status in real-time, which can be considered as **active** user feedback (Algorithm 1). Alternatively, the algorithm asks for more targeted user feedback according to specific selection principles so that the user can respond to those requests, which can be considered as **passive** user feedback (Algorithm 2).

The selection principles used in Algorithm 2 have a strong influence on the adaptability of the model to new subjects. A well-defined selection principle can largely accelerate the personalization and achieve a better detection performance. However, finding the points which provide the most information about the shift in distribution is non-trivial. We present three selection principles and their corresponding use cases. We assume that the user interacts with the device M times in total.

3.2.1. Random Selection Principle

The algorithm randomly picks M points from the test subject to form the new reference point list. It mimics the user actively interacting with the wearable device M times and provides the corresponding labels in real-time

3.2.2. Max Distance Selection Principle

The distances to the nearest reference points are computed for all incoming points and the M points with the largest distance are then picked. The idea is that the points that are far away from the reference points are more informative and can, therefore, reduce the uncertainty of the model.

3.2.3. Random Correction Selection Principle

In each step, the algorithm randomly selects a misclassified point from the test subject as a new reference point and updates the model. Misclassified points deliver more insights into the difference between training and target data distribution and are considered more important to address the

Algorithm 1 Personalization with active user feedback

Input: train input $\mathbf{X}$, train labels $\mathbf{y}$ and number of interactions M
Initialize the reference point matrix $\mathbf{R}$ with the mean points for all unique labels $\mathbf{y}_r$
Train a classifier $\mathcal{C}$ on $\mathbf{R}$ with labels $\mathbf{y}_r$
for $i = 1$ **to** M **do**
  Retrieve the current data point $\mathbf{x}$ and extract the label y from the interaction
  if the mean point $\boldsymbol{\mu}$ of label y still exists in $\mathbf{R}$ **then**
    Replace $\boldsymbol{\mu}$ with $\mathbf{x}$
  else
    Concatenate $\mathbf{x}$ to the end of $\mathbf{R}$, y to the end of $\mathbf{y}_r$
  end if
  Retrain $\mathcal{C}$ on new $\mathbf{R}$ and $\mathbf{y}_r$
end for

Algorithm 2 Personalization with passive user feedback

Input: train input $\mathbf{X}$, train labels $\mathbf{y}$, number of interactions M and selection criteria ϕ
Initialize the reference point matrix $\mathbf{R}$ with the mean points for all unique labels $\mathbf{y}_r$
Train a classifier $\mathcal{C}$ on $\mathbf{R}$ with labels $\mathbf{y}_r$
repeat
  Collect the data point $\mathbf{x}$ from the sensor in wearable devices
  if selection criteria $\phi(\mathbf{x}, \mathbf{R})$ is satisfied **then**
    Post a labeling request and extract the label y from the interaction
    if the mean point $\boldsymbol{\mu}$ of label y still exists in $\mathbf{R}$ **then**
      Replace $\boldsymbol{\mu}$ with $\mathbf{x}$
    else
      Concatenate $\mathbf{x}$ to the end of $\mathbf{R}$, y to the end of $\mathbf{y}_r$
    end if
  end if
  Retrain $\mathcal{C}$ on the new $\mathbf{R}$ with its labels
until c is equal to M

inter-individual variability. Essentially, this selection principle simulates the situation where users passively correct the prediction results presented by the wearable device.

4. Experiments

4.1. Datasets

In our experiments, we used two public datasets as well as our own "EIKEN" dataset for stress detection. All three datasets contain heartbeat-related data from which we extracted R-R intervals. We use this signal as a basis to examine and compare our approach. A short description of the public datasets is given in section 4.1.1, while details to the "EIKEN" dataset are provided in section 4.1.2.

4.1.1. Public Datasets

The WESAD (wearable stress and affect detection) dataset (Schmidt et al., 2018a) contains multi-modal physiological signals including blood volume pulse, electrocardiogram, electrodermal activity, electromyogram, respiration, body temperature, and three-axis acceleration. The signals were collected from 15 participants under three conditions: baseline, amusement, and stress. The stress trigger for the stress condition is the well-studied Trier social stress test (TSST) (Kirschbaum et al., 1993), which consists of a public speaking and a mental arithmetic task. The experimental procedure also included a preparation, meditation, and recovery phase with the purpose to reduce perturbations and acquire a clean affective condition.

The SWELL knowledge work (SWELL-KW) dataset (Koldijk et al., 2014) is a multi-modal dataset for work stress research. 25 participants performed typical office work under three different stress conditions: neutral condition, time pressure, and e-mail interruptions. Each stress condition experiment lasted around 30 minutes. The dataset contains electrocardiogram, skin conductance, computer logging, facial expression, and body posture. We use the raw heart rate data with three different class labels for our experiments.

4.1.2. EIKEN Dataset

The EIKEN dataset was acquired as part of the DeStress project that focused on building models to detect, distinguish, and classify physical (absolute stress), mental stress (relative stress), and combined mental and physical stress in real-time using heart rate variability (HRV) analysis and other biosignals using low-cost wearables. An initial effort to train models using HRV data collected in laboratory conditions resulted in a C5 decision tree with precision, recall, and F-score of close to 90% (Pluntke et al., 2019). The DeStress data acquisition system consists of two servers (server.ts and destress-server.py), an Android application, a TypeScript desktop application (firefighters-monitor-app), and a Mongo database. Multiple participants wearing a sensor were connected via an Android App to the server. Once the polar H10 chest band is paired with the Android app it posts reports to the server.ts which are stored in the Mongo database. The classifier is notified and starts classifying stress using the trained models from (Pluntke et al., 2019). The resulting HR and classification is displayed on the monitoring app. For combined stress, feedback was requested from the expert every 30 seconds or when there is a stress-level classification change.

The DeStress system was tested on 85 "instrumented smoke

divers" in the Civil Emergency Services Training Center at Eiken AG in Switzerland where 150 firefighters were certified as smoke diving team leaders in a darkened chamber with a 3D maze of cages simulating a building on fire. The firefighters needed to find their way and rescue unconscious people using only the illumination from phosphorescent helmets while being disoriented and stressed by smoke, strobing lights, and disturbing noise. Examiners evaluated the participants using infrared cameras on their time of transit, the ability to find all objects, and the quality of teamwork. This puts participants through 10-20 minutes of intense physical and mental stress (Gerke et al., 2019).

The participants were given chest belts (Polar H10) and "communication hub" smartphones 5-30 minutes before entering the maze. Afterward, feedback on their perception of the difficulty of the exercise was collected. The R-R intervals were transmitted from the Polar H10 via Bluetooth to the smartphone app and via Wifi to the Mongo database on the control system. Outliers in the R-R intervals were replaced with their nearest normal neighbors (i.e. Winsorized). C5 decision tree models predicted correct stress classes (Pluntke et al., 2019) but convolutional neural networks models provided a better classification of combined stress. Self-assessment was less reliable than assessment by a human expert, but no statistical analysis of expert feedback was possible because the expert was overloaded to have multiple firefighters under control. For labeling, we assumed and verified with occasional expert feedback that the presence in the cage maze resulted in mental stress and after 3 minutes in combined stress.

Analysis of the EIKEN dataset using K-means clustering on 18 engineered features for k = 2 did not show a clear separation when plotted in two dimensions using mean heart rate (MeanHR) and root mean square of successive differences (RMSSD), top biomarkers of stress (Blásquez et al., 2009; Sun et al., 2010). This was due to the high dimensionality as well as the tendency of K-means to cluster samples based on the Euclidean distance from cluster centroids, regardless of true separation. Convolutional autoencoders with their much fewer trainable weights produced clusters that were distinct with more than 90% of the samples mentally stressed and less than 10% with normal HRV. A shortcoming was that the method did not take into account intrinsic differences in HRV of different individuals (Oskooei et al., 2019).

For the subsequent analysis the 2-20 minutes waiting phase before entry into the cage maze was used to create a baseline. The first 20% of the time in the cage maze was labeled as mental stress while the further 80% of the time in the cage maze was labeled as combined stress. Verification of the recorded data showed that this is a good assumption. To eliminate labelling errors due to false timestamps for entry and exit, we checked whether the pulse in the cage maze was at least 10% higher than in the waiting phase and that the pulse during the first 20% of the cage maze was higher than in the waiting phase. After this pruning we analyzed data from 75 participants. A comparison of the three datasets is shown in Table 2:

Table 2. Dataset comparison

	WESAD	SWELL-KW	EIKEN
PARTICIPANTS	15	25	75
CLASSES	BASELINE, AMUSEMENT, STRESS	NEUTRAL, TIME PRESSURE, INTERRUPTION	DARKNESS, CLAUSTROPHOBIA, NOISE
STRESSORS	TSSI	EMAIL INTERRUPTIONS, TIME PRESSURE	TIME PRESSURE, DES-ORIENTATION
ECG SENSOR	RESPIBAN	MOBI	POLARH10

4.2. Results

In our first experiment we examined the feature distributions under different stress conditions. Since the dimension of the embedding space for our self-supervised model can be freely chosen, it is well suited for visualization. Figure 1 shows the learned features for three randomly selected test subjects and the same features after merging the individual data. The learned features show good separability of the different conditions for each test subject. However, once the data for the different subjects are combined, the classes partially overlap. For example, it can be observed that in the WESAD dataset, the baseline and amusement conditions are more difficult to separate after merging the data. The same applies to the two stress conditions in the SWELL and EIKEN datasets.

In the next experiment we investigated the approach to simply mix the data of the training subjects with additional calibration points from the test subject. The first three calibration points were randomly picked from each condition, so that at least one point per class was selected. The following points were then chosen randomly without considering their labels. After each new point we retrained the classifier and averaged the accuracies over all test subjects. Figure 2 shows the performance of this approach for different numbers of prior points taken from the $N-1$ training subjects for three different classifiers: Random forest (RF), linear discriminant analysis (LDA) and k nearest neighbour (k-NN) with $k = 1$. The black lines correspond to the purely person-specific approach. In this case, no points from other test subjects were used. The asymptotic left-most point of the purple line can be considered as the person-generic approach (with all points of the $N-1$ training participants and without calibration points). We can observe that, in this setup, the use of more points from the training subjects generally diminishes the performance. Moreover, in the case where all points from the $N-1$ training subjects were

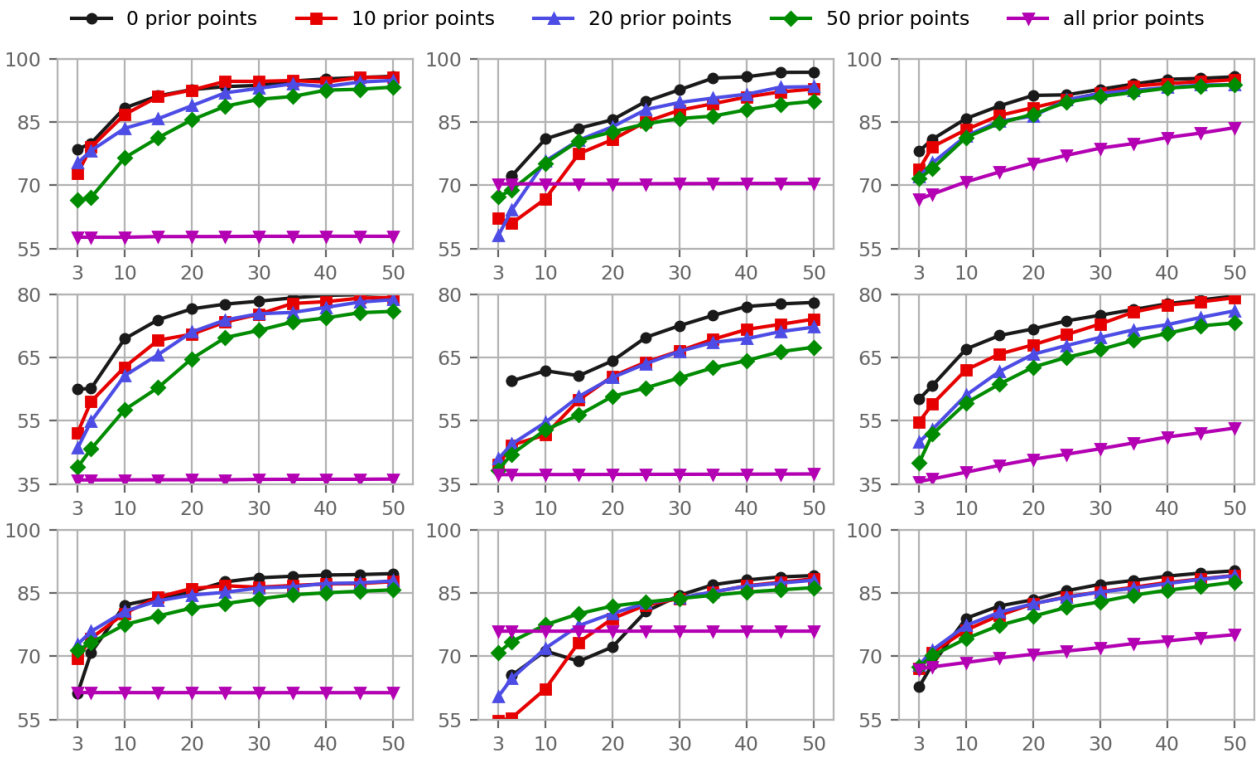

Figure 2. Performance of the mixing approach for different numbers of calibration points per subject on WESAD (top row), SWELL (center row) and EIKEN (bottom row) for the following classifiers: RF (left column), LDA (center column), and 1-NN (right column). Each curve corresponds to a different number of prior points taken from the training subjects. The accuracies are averaged over all test subjects

used, additional calibration points did not improve the accuracy for the RF and LDA classifier. This demonstrates the inter-individual differences in the feature distribution.

In our final experiment we compared different personalization approaches for different numbers of user interactions (or calibration points). For the baseline calibration approach, this number is equivalent to the number of segments used to compute the feature mean of the baseline condition. The results are shown in Figure 3. Each approach was evaluated based on HRV and SSL features. We observe the best performance with the k-NN classifier based on SSL features. While the baseline calibration approach shows good performance for little user interaction, it is outperformed by the other approaches when more feedback is provided. Furthermore, the personalization approach with active user feedback generally outperforms the approaches based on the maximum distance and random selection criteria.

5. Conclusion and Future Work

In this paper, we investigated different personalization strategies for stress detection with wearable devices and compared them with baselines on three datasets. Our analysis of the feature distributions demonstrate that each subject shows slightly different physiological responses to stress which is difficult to capture with a generic model. Combined with the observed diminished performance when including additional samples from the N$-$1 training subjects, this indicates the need for personalized methods. Our proposed calibration method addresses this problem from two perspectives:

1. The SSL features introduced avoid potential distortion from noisy labels, and a smaller feature dimension (d $= 5$) compared to HRV parameters (d $= 17$) helps mitigate the curse of dimensionality, so less interaction is required for personalization.

2. More informative user feedback is provided by the proposed interaction principles for the different scenarios.

After we finish this work, we are interested in three re-

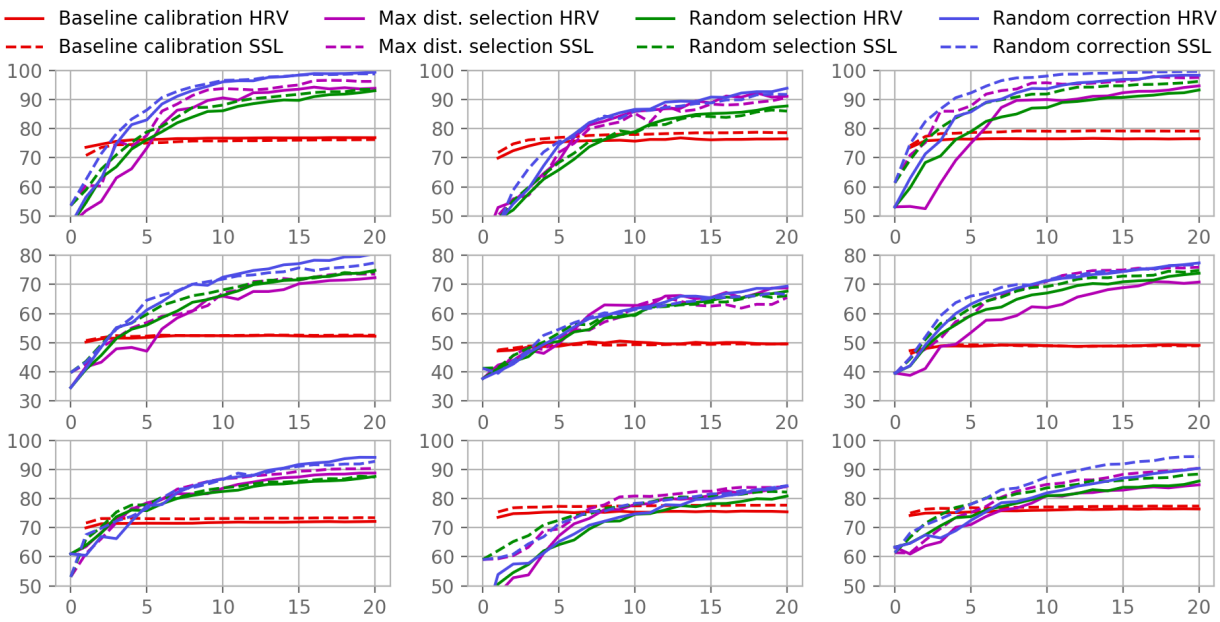

Figure 3. Performance comparison of different personalization approaches on the WESAD (top row), SWELL (center row) and EIKEN dataset (bottom row) for the following classifiers: random forest (left column), decision tree (center column), and 1-nearest neighbour classifier (right column). Each subplot shows the accuracy over the number of used calibration points. The accuracies are averaged over all subjects

search ideas. To better test our approach, we plan to acquire datasets with better personalized and standardized label acquisition for combined physical and mental stressors. To this end, test subjects will be exposed to mental exercises/games while cycling with defined physical loads on a bicycle home trainer (Sierro, 2020). As the human-in-the-loop approach showed its advantage in stress detection, we are going to investigate how to combine user modeling with interactive approaches such that less data is required to learn the relationship between user profile and inter-individual variability. Furthermore, we want to validate our model under field conditions and extend the input to other biosignals such as electroencephalogram, electrodermal activity, and electromyography.

6. Acknowledgement

This research was co-funded by the Bavarian Ministry of Economic Affairs, Regional Development and Energy, project Dependable AI, IBM Deutschland GmbH, and IBM Research, and was carried out within the Center for AI jointly founded by IBM and fortiss

References

Ayzenberg, Y., Hernandez Rivera, J., and Picard, R. Feel: frequent eda and event logging–a mobile social interaction stress monitoring system. In *CHI'12 extended abstracts on human factors in computing systems*, pp. 2357–2362. ACM, 2012.

Baddeley, A. D. Selective attention and performance in dangerous environments. *British journal of psychology*, 63(4):537–546, 1972.

Benson, H. and Allen, R. L. How much stress is too much? *Harvard Business Review*, 58(5):86–92, 1980.

Blásquez, J. C. C., Font, G. R., and Ortís, L. C. Heart-rate variability and precompetitive anxiety in swimmers. *Psicothema*, 21(4):531–536, 2009.

Bogomolov, A., Lepri, B., Ferron, M., Pianesi, F., and Pentland, A. Daily stress recognition from mobile phone data, weather conditions and individual traits. In *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 477–486, 2014.

Boonnithi, S. and Phongsuphap, S. Comparison of heart rate variability measures for mental stress detection. In *2011 Computing in Cardiology*, pp. 85–88. IEEE, 2011.

Can, Y. S., Arnrich, B., and Ersoy, C. Stress detection in daily life scenarios using smart phones and wearable sensors: A survey. *Journal of biomedical informatics*, pp. 103139, 2019.

Carneiro, D., Novais, P., Augusto, J. C., and Payne, N. New methods for stress assessment and monitoring at the workplace. *IEEE Transactions on Affective Computing*, 10(2):237–254, 2017.

Costin, R., Rotariu, C., and Pasarica, A. Mental stress detection using heart rate variability and morphologic variability of eeg signals. In *2012 International Conference and Exposition on Electrical and Power Engineering*, pp. 591–596. IEEE, 2012.

Dickerson, S. S. and Kemeny, M. E. Acute stressors and cortisol responses: a theoretical integration and synthesis of laboratory research. *Psychological bulletin*, 130(3): 355, 2004.

Doersch, C., Gupta, A., and Efros, A. A. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1422–1430, 2015.

Fails, J. A. and Olsen Jr, D. R. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*, pp. 39–45, 2003.

Gerke, S., Pluntke, U., Sridhar, A., Weiss, J., and Michel, B. Stresslevelbestimmung in der Atemschutzausbildung: Echtzeitanalyse von koerperlicher und mentaler Stressbelastung. *Brandschutz Deutsche Feuerwehr-Zeitung*, 2019 (3):178–184, 2019.

Ghaderi, A., Frounchi, J., and Farnam, A. Machine learning-based signal processing using physiological signals for stress detection. In *2015 22nd Iranian Conference on Biomedical Engineering (ICBME)*, pp. 93–98. IEEE, 2015.

Giakoumis, D., Tzovaras, D., and Hassapis, G. Subject-dependent biosignal features for increased accuracy in psychological stress detection. *International Journal of Human-Computer Studies*, 71(4):425–439, 2013.

Gidaris, S., Singh, P., and Komodakis, N. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

Gjoreski, M., Gjoreski, H., Luštrek, M., and Gams, M. Continuous stress detection using a wrist device: in laboratory and real life. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, pp. 1185–1193, 2016.

Healey, J. A. and Picard, R. W. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on intelligent transportation systems*, 6(2): 156–166, 2005.

Hobfoll, S. E. Conservation of resources: A new attempt at conceptualizing stress. *American psychologist*, 44(3): 513, 1989.

Horowitz, M. J. Stress response syndromes. *American Psychological Association*, 1976.

Huysmans, D., Smets, E., De Raedt, W., Van Hoof, C., Bogaerts, K., Van Diest, I., and Helic, D. Unsupervised learning for mental stress detection-exploration of self-organizing maps. *Proc. of Biosignals 2018*, 4:26–35, 2018.

Jing, L. and Tian, Y. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

Kirschbaum, C., Pirke, K.-M., and Hellhammer, D. H. The 'trier social stress test'–a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology*, 28(1-2):76–81, 1993.

Koldijk, S., Sappelli, M., Verberne, S., Neerincx, M. A., and Kraaij, W. The swell knowledge work dataset for stress and user modeling research. In *Proceedings of the 16th international conference on multimodal interaction*, pp. 291–298, 2014.

Koldijk, S., Neerincx, M. A., and Kraaij, W. Detecting work stress in offices by combining unobtrusive sensors. *IEEE Transactions on Affective Computing*, 9(2):227–239, 2016.

Kolodyazhniy, V., Kreibig, S. D., Gross, J. J., Roth, W. T., and Wilhelm, F. H. An affective computing approach to physiological emotion specificity: Toward subject-independent and stimulus-independent classification of film-induced emotions. *Psychophysiology*, 48(7):908–922, 2011.

Lupien, S. J., Maheu, F., Tu, M., Fiocco, A., and Schramek, T. E. The effects of stress and stress hormones on human cognition: Implications for the field of brain and cognition. *Brain and cognition*, 65(3):209–237, 2007.

Mason, J. W. A review of psychoendocrine research on the sympathetic-adrenal medullary system. *Psychosomatic medicine*, 30(5):631–653, 1968.

Melillo, P., Bracale, M., and Pecchia, L. Nonlinear heart rate variability features for real-life stress detection. case study: students under stress due to university examination. *Biomedical engineering online*, 10(1):96, 2011.

Mokhayeri, F., Akbarzadeh-T, M., and Toosizadeh, S. Mental stress detection using physiological signals based on soft computing techniques. In *2011 18th Iranian Conference of Biomedical Engineering (ICBME)*, pp. 232–237. IEEE, 2011.

Mozos, O. M., Sandulescu, V., Andrews, S., Ellis, D., Bellotto, N., Dobrescu, R., and Ferrandez, J. M. Stress detection using wearable physiological and sociometric sensors. *International journal of neural systems*, 27(02): 1650041, 2017.

Munla, N., Khalil, M., Shahin, A., and Mourad, A. Driver stress level detection using hrv analysis. In *2015 International Conference on Advances in Biomedical Engineering (ICABME)*, pp. 61–64. IEEE, 2015.

Nkurikiyeyezu, K., Yokokubo, A., and Lopez, G. The influence of person-specific biometrics in improving generic stress predictive models. *arXiv preprint arXiv:1910.01770*, 2019.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Oskooei, A., Chau, S. M., Weiss, J., Sridhar, A., Martínez, M. R., and Michel, B. Destress: Deep learning for unsupervised identification of mental stress in firefighters from heart-rate variability (hrv) data. *arXiv preprint arXiv:1911.13213*, 2019.

Picard, R. W., Vyzas, E., and Healey, J. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE transactions on pattern analysis and machine intelligence*, 23(10):1175–1191, 2001.

Pluntke, U., Gerke, S., Sridhar, A., Weiss, J., and Michel, B. Evaluation and classification of physical and psychological stress in firefighters using heart rate variability. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 2207–2212. IEEE, 2019.

Poria, S., Cambria, E., Bajpai, R., and Hussain, A. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125, 2017.

Reiss, A. and Stricker, D. Personalized mobile physical activity recognition. In *Proceedings of the 2013 international symposium on wearable computers*, pp. 25–28, 2013.

Rodrigues, S., Paiva, J. S., Dias, D., and Cunha, J. P. S. Stress among on-duty firefighters: an ambulatory assessment study. *PeerJ*, 6:e5967, 2018.

Schmidt, P., Reiss, A., Duerichen, R., Marberger, C., and Van Laerhoven, K. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pp. 400–408, 2018a.

Schmidt, P., Reiss, A., Duerichen, R., and Van Laerhoven, K. Wearable affect and stress recognition: A review. *arXiv preprint arXiv:1811.08854*, 2018b.

Setz, C., Arnrich, B., Schumm, J., La Marca, R., Tröster, G., and Ehlert, U. Discriminating stress from cognitive load using a wearable eda device. *IEEE Transactions on information technology in biomedicine*, 14(2):410–417, 2009.

Sharma, N. and Gedeon, T. Stress classification for gender bias in reading. In *International Conference on Neural Information Processing*, pp. 348–355. Springer, 2011.

Shi, Y., Nguyen, M. H., Blitz, P., French, B., Fisk, S., De la Torre, F., Smailagic, A., Siewiorek, D. P., al'Absi, M., Ertin, E., et al. Personalized stress detection from physiological measurements. In *International symposium on quality of life technology*, pp. 28–29, 2010.

Sierro, N. Firefighter vital sign monitoring for predicting operational readiness. *EPFL Master thesis*, 2020.

Srinivas, A., Laskin, M., and Abbeel, P. Curl: Contrastive unsupervised representations for reinforcement learning. *arXiv preprint arXiv:2004.04136*, 2020.

Sun, F.-T., Kuo, C., Cheng, H.-T., Buthpitiya, S., Collins, P., and Griss, M. Activity-aware mental stress detection using physiological sensors. In *International conference on Mobile computing, applications, and services*, pp. 282–301. Springer, 2010.

Sztyler, T. and Stuckenschmidt, H. Online personalization of cross-subjects based activity recognition models on wearable devices. In *2017 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pp. 180–189. IEEE, 2017.

Valenza, G., Citi, L., Lanatá, A., Scilingo, E. P., and Barbieri, R. Revealing real-time emotional responses: a personalized assessment based on heartbeat dynamics. *Scientific reports*, 4:4998, 2014.

Vanitha, V. and Krishnan, P. Real time stress detection system based on eeg signals. *Biomedical Research*, 2016.

Vidulich, M. A., Stratton, M., Crabtree, M., and Wilson, G. Performance-based and physiological measures of situational awareness. *Aviation, space, and environmental medicine*, 1994.

Wijsman, J., Grundlehner, B., Penders, J., and Hermens, H. Trapezius muscle emg as predictor of mental stress. In *Wireless Health 2010*, pp. 155–163. ACM, 2010.

Xu, H. and Mannor, S. Robustness and generalization. *Machine learning*, 86(3):391–423, 2012.

Zenonos, A., Khan, A., Kalogridis, G., Vatsikas, S., Lewis, T., and Sooriyabandara, M. Healthyoffice: Mood recognition at work using smartphones and wearable sensors. In *2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*, pp. 1–6. IEEE, 2016.

Zhai, J. and Barreto, A. Stress detection in computer users based on digital signal processing of noninvasive physiological variables. In *2006 international conference of the IEEE engineering in medicine and biology society*, pp. 1355–1358. IEEE, 2006.

Zhang, R., Isola, P., and Efros, A. A. Colorful image colorization. In *European conference on computer vision*, pp. 649–666. Springer, 2016.