Insightful Stress Detection from Physiology Modalities using Learning Vector Quantization

J.J.G. (Gert-Jan) de Vries[a,b,*], Steffen C. Pauws[a], Michael Biehl[b]

[a] *Philips Research - Healthcare, High Tech Campus 34, 5656 AE Eindhoven, The Netherlands*

[b] *Johann Bernoulli Institute for Mathematics and Computer Science, University of Groningen, The Netherlands*

Abstract

Stress in daily life can lead to severe conditions as burn-out and depression and has a major impact on society. Being able to measure mental stress reliably opens up the ability to intervene in an early stage. We performed a large-scale study in which skin conductance, respiration and electrocardiogram were measured in semi-controlled conditions. Using Learning Vector Quantization techniques, we obtained up to 88% accuracy in the classification task to separate stress from relaxation. Relevance learning was used to identify the most informative features, indicating that most information is embedded in the cardiac signals. In addition to commonly used features, we also explored various novel features, of which the very-high frequency band of the power spectrum was found to be a very relevant addition.

Keywords: mental stress, classification, Learning Vector Quantization

1. Introduction

Daily life becomes increasingly stressful. While certain levels of psychological stress help us perform optimally, prolonged exposure to stressors can have severe effects on wellbeing. Chronic stress is known to contribute to the development of, among others, cardiovascular diseases [1, 2] and has been found to contribute to high societal costs. In the US, for example,

[*] Corresponding author. tel. +31 6 52773212; fax. +31 40 2746321; High Tech Campus 34, 5656 AE Eindhoven, The Netherlands

Email address: gj.de.vries@philips.com (J.J.G. (Gert-Jan) de Vries)

it has been estimated that job stress costs ”over $300 billion annually due to increased absenteeism, employee turnover, diminished productivity, medical, legal and insurance expenses, and workers’ compensation payments” [3].

The fine balance between the positive effects of short term stress and the detrimental effects of chronic stress on the one hand, and an increasingly demanding society on the other hand, indicate the need for assistance in balancing workload. Various products are available to help regulate mental stress [4, 5] including various biofeedback systems. One such biofeedback method is the stimulation of alpha-frequency brain waves, i.e., alpha neurofeedback [6]. Alpha brain waves are related to relaxation during wake, and stimulation of these waves are known to increase relaxation levels [7]. The effects have been studied in the lab quite extensively, but only limitedly in circumstances that better reflect daily life. The application of neuro-feedback in a consumer device using the paradigm of music listening was researched[8] in a double-blinded experiment with two types of control, as one aim of a comprehensive study. The effectiveness of such methods can, however, be further improved by providing them at the right moment to the right people. To that end, an objective method of measuring stress using easily and unobtrusively measurable physiological parameters is needed. This lead to the second aim of the aforementioned study: the development of such a method; which is subject of the present manuscript.

Several studies have attempted to classify stress from physiological measurements [9, 10, 11, 12, 13, 14] using various classification techniques. Among the more popular are Support Vector Machine (SVM) and Artificial Neural Network (ANN) [15]. Learning Vector Quantization (LVQ) is a relatively novel technique that has been applied successfully to a wide range of classification challenges [16], but rarely to classification of affect, and to the best of our knowledge, not yet to stress classification. The family of LVQ classification techniques use prototypes that are defined in the same mathematical space as the input data. The intuitive nature and ease of inspection give LVQ an advantage over less open-box methods such as SVM and ANN. We exploit this property of LVQ to gain new insights in the field of mental stress detection, where further understanding of the domain can help improve descriptive models[15].

In the present study, we set out to build classifiers to distinguish stress from relaxation using the three modalities of Electrocardiogram (ECG), Galvanic Skin Response (GSR), and Respiration (RSP). Moreover, we set out to explore the application of LVQ methods in this domain. As a reference method we also apply SVM. We will use these methods to explore their performance as affective classifiers, to compare uni-modal and multimodal classifiers in order to find out which signal is most rich in information to distinguish stressful reactions and to investigate how individual features contribute. In the following, we will first create an overview of published affective and stress classifiers, then we describe the methods used, followed by results, discussion and conclusion.

2. Affect and Stress Classification

Whereas there are multiple definitions of stress that differ in various subtleties, an often used definition is that of Lazarus & Folkman: "Psychological stress is a relationship between the person and the environment that is appraised by the person as taxing or exceeding his or her resources and endangering his or her well-being" [17]. Stress can be measured through a variety of physiological signals, among which Skin Conductance (SC), Skin Temperature (ST), ECG, Blood Volume Pulse (BVP), Blood Pressure (BP), Electroencephalogram (EEG) and Electromyogram (EMG) [15]. Because emotions and other affective states can also be measured using these signals, it is worth positioning our work in the light of other affective classifications as well.

Table 1 shows shows a snapshot of ten affect classification studies from physiology. It can be seen that a variety of physiological modalities is used as input, various techniques are applied and a variety of target classes are used. Because the number of classes, number of participants, prior probability of classes and methods used for validation vary between these studies, their performance cannot be compared directly. Nevertheless one can observe that there is room for improvement in terms of performance, which ranges between 61% and 86%, with the majority of performances between 70 and 80%.

Table 2 shows a detailed overview of studies that specifically classify mental stress. We observe that the performances reported are slightly higher than those reported for other affective states (Table 1). We observe that most studies report 'ordinary' cross validation in which data of participants is shared over training and test set, only a limited number of studies report participant-wise cross validation results in which participants are strictly separated over training and test set (i.e., no data of test-participants is used for training). The latter is generally more difficult than the former, which becomes also apparent in the performances in Table 2, but does better reflect the generalization performance (i.e., performance of the method for unseen users).

The study of Healey & Picard [9] provided an exceptionally high performance of 97%. It should, however, be noted that their study is limited in the number of participants used (13) as well as using only one task for each stress level. Therefore, the high performance they obtained is likely biased by the specific set of participants and might reflect distinctions between the tasks rather than the stress levels. In general, we observe that the number of participants used in the studies is relatively limited: the studies included data from 3 to 32 participants. In our study we gathered data from more participants to have a more representative set of participants. We repeated measurements in 15 sessions to introduce temporal effects and environmental changes in the dataset that happen in daily life and influence the physiological measurements. Furthermore, we use multiple stressful tasks to induce more variety to better represent stressful situations in daily life.

Sharma & Gedeon [15] made an extensive inventory of various aspects of stress detection. They conclude that "Models developed to date that describe stress are quite simplistic. Generally, established techniques such as ANN and SVM have been used to model stress. Novel or more complex computational techniques are needed for stress models". We believe that the application of LVQ classifiers can be such a novel computational technique and help gain more direct insight into the stress classification challenge and thereby provide valuable input to develop models that describe stress.

3. Method

The experiment performed to obtain the data that will be used in the analysis that is subject of this work is further described in [8]. The following sections describe the most important details. The current problem is defined as a binary classification problem in discerning stressful from relaxation episodes from human physiological signals. Stressful episodes were operationalized as various mentally demanding tasks, relaxation episodes were operationalised listening to favourite music. Human physiological signals entail the following modalities: ECG, GSR and RSP.

3.1. Participants

Participants were recruited by means of a website that explained the procedures involved in the research in great detail. A total number of 171 persons indicated on the website that they wanted to participate in the research. 110 persons either did not follow up on our request, turned out to be unavailable at the time of the research, or decided to cancel their participation. The remaining 61 (20 male, 41 female) provided written informed consent. Their age ranged from 18 to 28 years (mean 21.2 years).

3.2. Design and procedure

Each participant returned 15 times within a period of 4 weeks for a session during which their physiology was measured. The sessions took place in a normal office room, in which each participant was seated in a comfortable reclining chair in front of a small table with a laptop on it. There were five such chairs and tables with laptops in the room, separated by wooden partitions, so that 5 participants could be trained at the same time by a single experimenter. The whole session was automated as much as possible. The experimenter supervised the sessions, and only took action in case something was wrong (usually bad electrode contacts, which were automatically signalled).

A training session on a particular day always consisted of the same sequence of tasks. After the signals were determined to be valid, a baseline measurement of five minutes rest with eyes opened was recorded, followed by 5 minutes with eyes closed. After that, 3 relaxation intervals of 8 minutes duration were interspersed by cognitive tasks lasting about 5 minutes each. The sequence of tasks are graphically represented in Figure 1. The (fixed) sequence was: Flanker task, relaxation 1, Stop-signal task, relaxation 2, Stroop task, relaxation 3, N-back task. During the relaxation intervals subconscious neuro-feedback was provided in three different ways, two of which are control

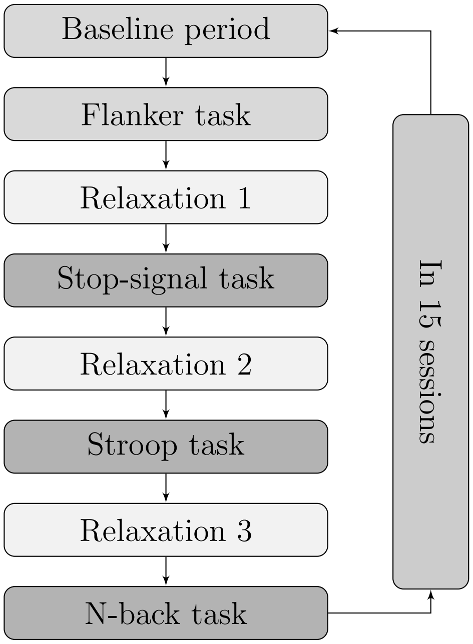

Figure 1: Schematic outline of the experiment

conditions.

The interleaved task sequence was chosen for several reasons. First, it represents daily life stress, secondly it enhances changes in stress level which are particularly of interest for practical applications, and thirdly it provides a platform to test the relaxation effect of neuro-feedback.

3.2.1. Relaxation with Neuro-feedback

The participants were given a set of headphones that they used for listening to their favorite music. Participants could either bring their own music for that particular day on an MP3 player, or they could select that day's music from a playlist containing thousands of songs from various artists. There was no limitation to the kind of music participants could listen to. Categories included genres like hard rock, easy listening and classical music.

As one part of this comprehensive study, the effects of neuro-feedback on relaxation were studied [8]. To that end, three conditions were used: alpha training and two types of controls, where one applies the same stimulation but at different (beta) frequencies that are not associated with relaxation and another control type where no stimulation is performed. Note that the stimulation was performed in a very subtle manner, as is described in the next paragraph, uses exactly the same setup over the three conditions, and has no direct effect on the peripheral physiological measurements taken (see Section 3.3).

The participants were randomly assigned to one of three groups: alpha training (A), random beta training (B), or control (C, music only); which was used for all sessions for this participant. Participants in group C listened to unaltered music, the music for the other two groups was altered by a high-pass filter of which the cut-off frequency was dynamically chosen. The cut-off frequency was adapted at real time based upon the frequency spectrum of the participants' EEG. To that end, the power in a target frequency range is calculated as relative to the total power (i.e., the power in the range 4-35Hz). The higher this relative power, the lower the cut-off frequency was chosen. The resulting effect is that lower (relative) power in the target frequency bands causes the low frequencies of the music to be filtered out, while high (relative) power in the target frequency will pass the music without much change (in the lower music frequencies). The target frequency ranges in the EEG spectrum were chosen as follows: The range for group A was based upon the power of alpha waves (8-12Hz), and for group B it was based upon beta waves (a randomized 4Hz bin in the range 16-36Hz). The alpha training was expected to increase relaxation while the other two types were not expected to have any effect on relaxation. These expectations were confirmed by Van Boxtel et al. [8].

From the 61 participants, 50 completed all training sessions (and without technical problems). The participants were distributed over the three groups as follows: A (alpha training): $N = 18$ (12 female; mean age 20.7 ± 1.8 years); B (random beta training): $N = 12$ (9 female; mean age 20.6 ± 1.5 years); C (control, music only): $N = 20$ (15 female; mean age 21.0 ± 2.1 years). Further details on the neuro-feedback training can be found in [8], the present study focusses at the stress and relaxation aspects of this study.

The mentally demanding tasks are further detailed in the following, taken from the study protocol [28].

3.2.2. Stop-signal task

” The stop-signal task basic choice reaction time task. A green triangle (0.050 of screen width) on a black background is presented on the computer screen. Subjects have to indicate as a fast as possible the direction of the triangle. For a triangle to the left, subjects press the most left button of a button box and when it points to the right, the most right button has to be pressed. In one third of the trials the green arrow becomes red for 100ms and no answer has to be given, as depicted in Figure 2. When subjects are able to stop their response, the next time the stop signal will be given 50ms later to make it more difficult. When subjects give a response despite the presence of a stop signal, the signal appears the next trial 50 ms earlier to make it easier for the subject to stop the response. The task starts with a stop signal delay time of 250 ms and depending on the reaction of the subject, the stop signal delay time changes. Logan and colleagues [29] fitted performance on this task in a formal model. The present task will use staircase tracking of response rate to arrive at 50% of successfully stopped trials, which is an optimal value for estimating inhibitory efficiency (Stop Signal Reaction Time (SSRT)). After one trial was finished, a fixation cross of 0.004 of the screen width appeared between 1 and 2 seconds on the screen before the next trial started. ” [28]

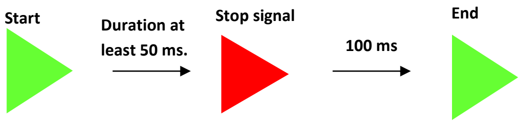

Figure 2: Example of a Stop-signal task. The trial starts with a green arrow that depending on the subject’s performance is green for a certain amount of time (at least 50 ms). After this time, the stop signal is initiated and the arrow becomes red for 100ms. This is followed by a green arrow that marks the end of the trial. It is the aim of the task that subjects do not give a response when the arrow becomes red.

3.2.3. Stroop task

” The computerized version of the Stroop Color Word Test (SCWT) [30] is used as a measure of executive functioning. In the Stroop task, subjects have to indicate whether the meaning of a word is the same as the color of which another word is printed in. Both words are not presented at exactly the same time to make it more difficult for the subject. In our version of the Stroop task, both words are printed above each other and the first word is presented 150 ms before the other word. In the case of Figure 3, the word "geel" is presented 150 ms before the word "rood". Both words are visible during 500 ms. In this period, subjects have to indicate whether the color of the upper word is the same as the meaning of the lower word. This requires inhibition of the automatic response to read the color word [31]. Hence, this test is considered a measure of 'disinhibition' and it generally has high reliability [32].

In the example, the color of the word "geel" is red and the meaning of the lower word is red, so the trial is correct. When the color and color name correspond subjects press with their index finger the 'yes'-button, if not, they press the 'no'-button. Whether the right or left index finger will be used for the yes and no response is counterbalanced between sessions. Congruent trials are trials in which the color of the upper word is the same as the meaning of this word. For example, the word "Blue" is written in blue ink and it means blue. When the color of the upper word is not the same as its meaning, the trial is called incongruent. In our experiment, four colors and the corresponding color names are used, namely red, yellow, blue and green. However, also the sign "XXXX" is used as an upper word. The expectation is that subjects will make fewer mistakes when "XXXX" is used as the upper word, because this word has no meaning and therefore subjects only have to deal with the color and not the meaning of the word. To keep between trials the attention of the subjects, a fixation cross with a variable duration between 1 and 2 seconds is presented on the computer screen. The duration of the fixation cross is variable to prevent a fixed rhythm of predicting and answering to the stimulus. " [28]

3.2.4. N-back task

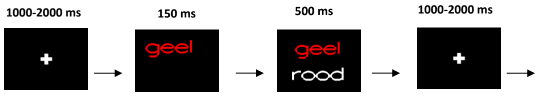

Figure 3: An example of an incongruent matching trial in the Stroop task. The word ”geel” means yellow but is written in red ink, so it is incongruent. The upper word is presented 150 ms before the lower word. Both words are visible during 500 ms. Between trials, a fixation cross with a duration between 1 and 2 seconds is presented on the computer screen.

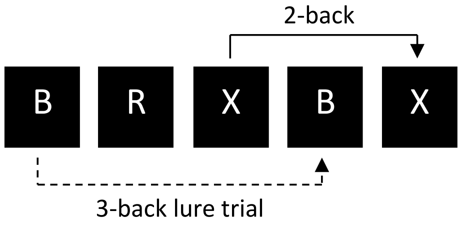

Figure 4: An example of a sequence of letters in the N-back task. The last X matches the letter that was presented two items ago (X) and is therefore a 2-back. In this case subjects have to press the ”yes”-button on the button box, indicating that it was a 2-back. The letter B on the fourth position of the sequence is the same as the first one and is an example of a 3-back lure trial.

” The N-Back task is a working memory task, introduced by Kirchner [33], and requires subjects to decide whether each stimulus in a sequence matches the one that appeared N items previously. For example in a 3-back task subjects have to decide whether a letter currently presented on the screen is the same as three letters earlier. Our version of the N-back task is a 2-back task, meaning that subjects should decide whether the letter on the screen was the same as two letters ago. The used test set consists of 8 letters, namely B, F, K, H, M, Q, R and X. We decided not to use vowels to prevent the formation of words, which are more easily remembered than single letters. Furthermore, we use letters who are spatially different to be sure that when subjects make an error, it is caused by the difficulty of the task and not by confusion whether a letter was a V or W for example. Each letter will be presented for half a second on the computer screen.

> Between the end of a stimulus and the beginning of the next one, a fixation cross appeared on the screen. Except the 2-back trials, also lure trials were included in the task, such as depicted in Figure 4. These were trials in which the trial was 1-back or 3-back. When a trial was 2-back, subjects respond to the target by pressing the 'yes' button with their index finger. Whether the right or left index finger has to be used for the yes response is counterbalanced between sessions. " [28]

3.2.5. Flanker task

> " The Eriksen Flanker task [34] is a basic choice reaction time task. In the task, five horizontally aligned arrows (size arrows: 0.050 of the screen width, space between arrows: 0.050 of the screen width) are presented on a 15 inch computer screen (resolution 1440 x 900 pixels, refresh rate 60 Hz) and subjects have to indicate the direction of the middle arrow (see Figure 5). Subjects can indicate this direction with a button box of which the most left button is pressed for an arrow pointing to the left and the most right button for an arrow pointing to the right. The two arrows on the left and right side of the middle arrow are flanker arrows and presented 150 ms before the middle arrow. These arrows are meant to distract the subject. The four flanker arrows always point in the same direction to the left or right. In this way, two situations can occur, namely that the flankers point in the same direction as the middle arrow (congruent) or that the flankers point in the opposite direction of the middle arrow (incongruent). The middle arrow with flankers will be present for 500 ms. After this period, a fixation cross (0.004 of screen width) appears at the same position as the middle arrow, namely in the center of the screen. To prevent that subjects learn when the next trial will start, the duration of the fixation cross will vary between 1 and 2 seconds. " [28]

For the classification analysis described in the Section 3.4, we selected the data gathered during the three relaxation tasks and the Stop-signal, Stroop and N-back tasks (as mentally stressful tasks). We did not include the Flanker task in the analysis as it turned out that participants were able

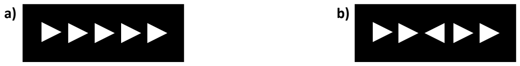

Figure 5: Stimuli used in the Flanker task. The two arrows on the left and right side of the middle arrow are the flanker arrows and presented 150 ms before the middle arrow appears. The arrows are white and presented on a black background. Subjects have to indicate the direction of the middle arrow. a) Congruent situation. The flanker arrows point in the same direction as the middle arrow. b) Incongruent situation. The flanker arrows point into the opposite direction of the middle arrow.

to master the Flanker task very well after only a few attempts, thereby strongly reducing the mental stressfulness of the task in subsequent sessions. After each task the participants were asked to rate their level of stress vs relaxation on a visual analogue scale. Effectiveness of the induction of stress vs relaxation was tested by applying an ANOVA with repeated measures to these reported levels of stress.

3.3. Measurements

GSR was recorded from the left index finger, ECG was recorded from an electrode placed on the left wrist, and RSP was measured using a chest belt with stretch sensor. The signals were sampled at a rate of 1024 Hz (ECG), and 256 Hz (GSR, and RSP) by a 24 bit A/D converter on a Nexus-10 portable device (MindMedia B.V., The Netherlands).

For each participant there were 15 sessions scheduled totalling 915 ($= 61 * 15$) sessions, of which 46 were discontinued due to technical problems or unconformity of participants, yielding 869 sessions. Due to bad signal quality (e.g., signals out of range of the measuring equipment) we further excluded, 96, and 121 sessions for GSR and RSP respectively, resulting in 772, and 748 sessions from analyses for these signals. No ECG sessions needed to be excluded. In total 662 sessions contained valid signals for all modalities. Figure 6 depicts the data selection (or exclusion) schematically.

3.3.1. Preprocessing & Feature extraction

The steps taken during preprocessing and feature extraction are schematically depicted in Figure 7. As a first step of preprocessing, the signals were

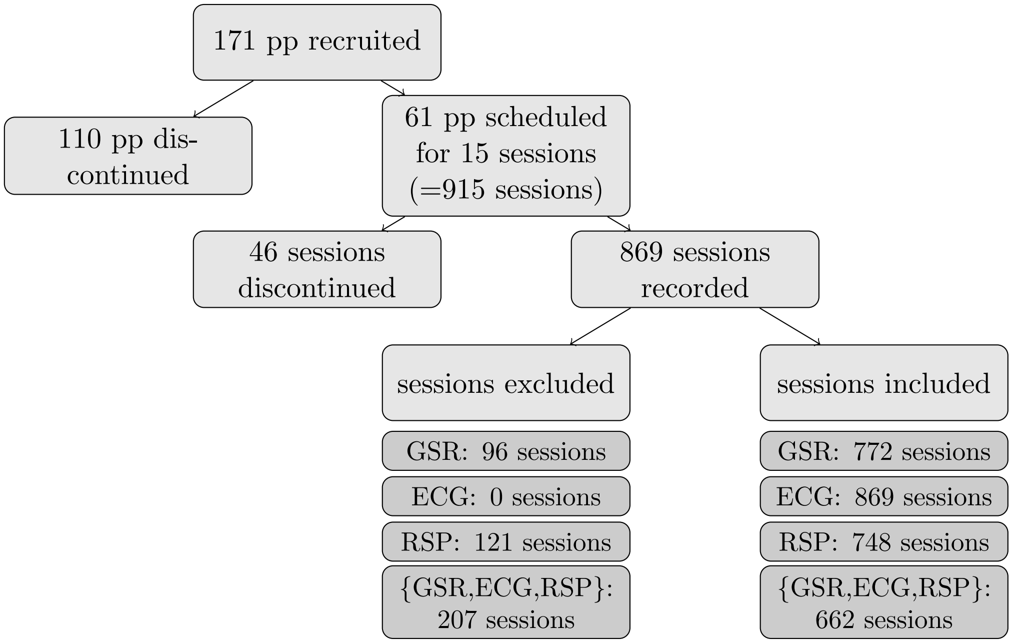

Figure 6: Schematic outline of the data selection/exclusion process. Left branches show exclusion, and right branches inclusion of sessions.

downsampled to 512 Hz (ECG), and 128 Hz (GSR, and RSP). Subsequently, signals were analyzed through the following dedicated preprocessing methods:

ECG preprocessing consisted of the following steps (as outlined in [35]): R-peak detection, IBI outlier removal, and Heart Rate Variability (HRV) analysis. R-peak detection was performed using a pattern matching technique [36]. The resulting intervals between the R-peaks, called Inter-Beat Intervals (IBI), are filtered for outliers by using a sliding window histogram. In order to estimate frequency domain HRV features, an Autoregressive-Moving Average (ARMA) time series model was used to derive power in the frequency bands defined in the HRV guidelines paper [37], ranging from 0.04 to 0.15 Hz, which is known to vary with parasympathetic nervous system activity [38]. Next to the frequency domain HRV features, a variety of time domain HRV features is calculated, given that no 'golden standard' for HRV has been defined [39], as well as several features based on plain IBIs.

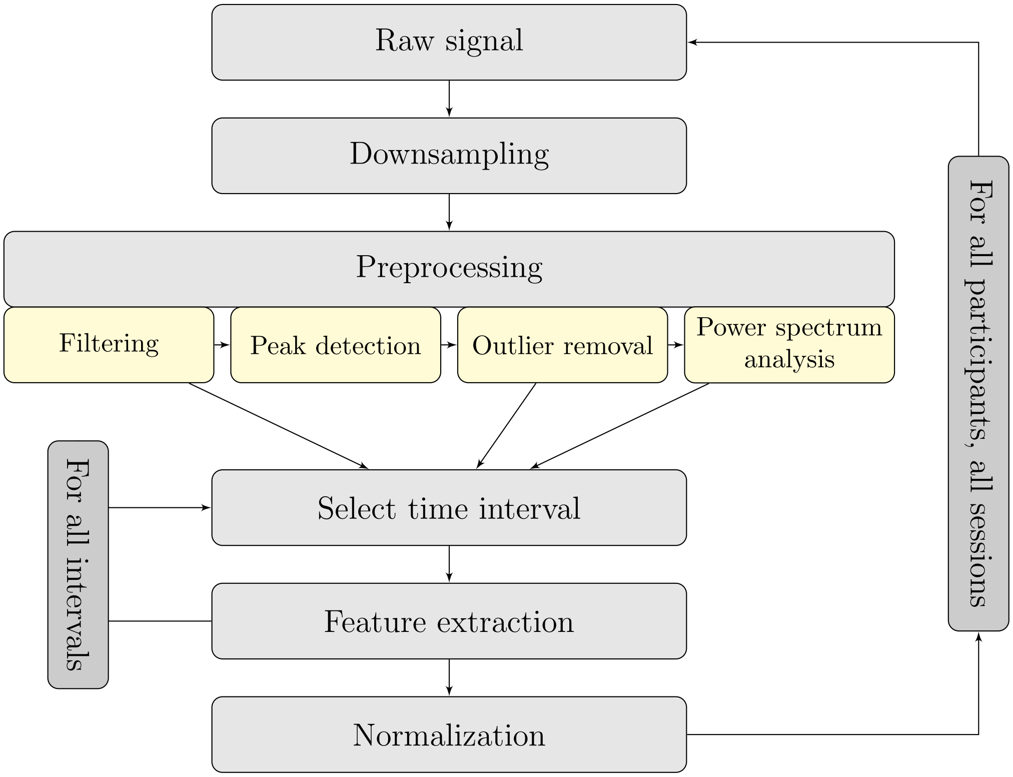

Figure 7: Schematic overview of the feature extraction process

GSR was preprocessed using the SCRGauge method described in [40] which first subsamples the GSR signal to 1 Hz, uses cubic splines interpolation followed by a dedicated local maximum detection which is triggered by exceeding a certain gradient. Backward and forward searches are subsequently applied to detect the onset of Skin Conductance Responses (SCRs), and half recovery times. The raw GSR signal was used to derive several Skin Conductance Level (SCL) features from, the extracted SCRs to derive SCR features from, and from the residual signal that resides after subtracting SCRs from the raw signal, using the technique described in [41] we derived features that represent purely the tonic part of GSR.

RSP signals were first lowpass filtered (cut-off 0.5Hz) and then analyzed for individual breaths. Using a localized min/max filter [42], local minima and maxima are detected. When found in the right order, they characterize a single breath. Based upon the distribution of identified breath amplitudes in a signal, too small or too large breaths (outliers) are removed. After this preprocessing the RSP signal is characterized by a sequence of breaths similar to the IBI signal for ECG.

All features have been calculated over equal length time intervals in order to avoid bias in duration dependent features (such as standard deviations) towards certain tasks. To this end, the first 5 minutes (which is the minimal duration of tasks) of measured signals from each task was taken to derived the feature values. A complete overview of extracted features can be found in Table 3. The specific features have been chosen such that they express the dynamics known to be relevant [43, 44, 45] as they are modulated by the Autonomous Nervous System (ANS) that responds to stress. From this large set of features, we compiled a subset of features representing the most often used features in literature (inspired by the list in [46]). They are marked with an asterisk in Table 3. In order to combine the data gathered from the different physiological signals to be used by a single classifier, we applied feature level fusion.

As highlighted in [47], there are many different techniques for normalization. In addition to the choice of which correction formula to use, the choice in defining the baseline period, there is also the choice of correcting on signal or feature level. The aim of normalization is to reduce the variance that occurs due to differences in physiology between participants, but also the long-term changes in physiology over time within participants, e.g., due to differences in physical fitness or the environment (such as temperature and humidity) [48, 49]. We have chosen for z-correction, a technique that compensates both for baseline level and variation, and is not too sensitive to outliers. Rather than applying the correction to the raw signals (which would only make sense for the skin conductance level), we apply it to the features derived, and we use the entire recording (all tasks) as reference signal, as suggested for e.g., SCR amplitude in [48, 49]. Hence, after computing the features per task, we applied z-correction ($x_{corr} = \frac{x-\mu}{\sigma}$) to compensate for differences in physiological baselines between people and sessions. In this formula, μ represents the mean of a feature's values over all tasks within a single session (for a single participant), and σ the respective standard deviation.

3.4. Classification analysis

In order to answer our research question of discerning stress from relaxation using physiology as input, we applied a selection of classifiers and further optimized their parameter settings using data from the individual physiological modalities (GSR, ECG and RSP) as well as the combined multimodal datasets consisting of pairwise combinations of modalities as well as all modalities combined. Finally, we used the trained LVQ classifiers to derive which features were most influential in distinguishing stress from relaxation.

Learning Vector Quantization (LVQ) comprises a family of classifiers that is of open box nature, that is, they provide direct insight into the information learned by the classifier. LVQ, initially proposed by Kohonen [50], defines prototypes $\mathbf{w}_T \in \mathbb{R}^N$ in the same (mathematical) space as the data (samples $\xi \in \mathbb{R}^N$) to represent the classes. These prototypes are directly interpretable as they show characteristics of classes in terms of the features chosen. During training, samples are presented sequentially, and for each sample the closest prototype(s) are updated by moving them towards or away from the presented sample. Several variants have been proposed, amongst which Robust Soft Learning Vector Quantization (RSLVQ) [51], which introduces soft prototype assignments which act similarly to a soft window around the decision boundary [52], and Generalized Matrix Learning Vector Quantization (GMLVQ) [53], which introduces relevance learning. Where typically in LVQ squared Euclidian distance, i.e., $d(\xi, \mathbf{w}_T) = (\xi - \mathbf{w}_T)^\top(\xi - \mathbf{w}_T)$, is used to measure distance between a data sample ξ and a prototype $\mathbf{w}_T$ indexed by T, GMLVQ uses pairwise weighing of the distance components through the following distance measure: $d_\Lambda(\xi, \mathbf{w}_T) = (\xi - \mathbf{w}_T)^\top\Lambda(\xi - \mathbf{w}_T)$. The relevance matrix $\Lambda \in \mathbb{R}^{N\times N}$ is trained along with the prototypes during the training phase and can be interpreted as a relevance matrix that indicates per pairwise combination of input dimensions how relevant they are to the classifier. To allow for regularization through the parameter $M \leq N$, we define $\Lambda = \Omega^\top\Omega$, with $\Omega \in \mathbb{R}^{M\times N}$. For further details on the GMLVQ algorithm, we refer to Schneider et al. [53].

We have trained GMLVQ both with $2 \times N$ and $5 \times N$ sized relevance matrices Ω, but since we observed identical performances, we will only report results of $2\times N$ sized Ω. We will present results for RSLVQ and GMLVQ using one prototype per class as using more prototypes per class did not improve the results. In addition, we apply SVM [54], a very popular technique in this domain of biomedical engineering. Next to linear SVM (of type C-SVM), which will be reported in the results, we also applied SVM with an Radial Basis Function (RBF) kernel, which however, did not improve upon the results. Hyper-parameters were optimized per method per parameter set using grid search.

3.4.1. Cross validation

In order to estimate generalization performance, we employed a cross validation scheme. Because physiological data shows large variation between participants [55, 48, 49], the most applicable, but also most challenging classification task is to separate training and test data not only per sample, but per subject. Hence we used 10 fold participant-wise cross validation, which divides the set of participants tenths and repeatedly uses data from 90% of participants for training and the rest for testing. The results reported are means and standard deviations over 10 × 10-fold participant-wise cross validations. In addition to the participant-wise cross validation, we also performed 'ordinary' cross validation in which data of single participants can be split over both training and test set, thereby leaking some information from 'training participants' to the test set.

4. Results

Per participant, we have 15 sessions comprising 3 repetitions of 2 tasks, one operating in a stress condition and one operating in a relaxation condition. As dependent variable, we asked participants to report stress level using a visual analogue scale from zero to one (mean for relax condition: 0.29; for stressful condition: 0.38). It is evident that mean reported stress levels are derived from same participants measured in different sessions, repetitions and tasks, not from different participants. This refers to a within-subject or repeated measure design for statistical analysis. To demonstrate the induction of stress by means of mental and relaxation tasks, the aim of the analysis is to reject the null hypothesis of no difference in mean 'reported stress level' between stress conditioned tasks and relaxation conditioned tasks. We conducted an ANOVA with repeated measures with 'reported stress level' as dependent variable, and session (15), repetitions (3) and tasks (2) as within-subject independent variables. Missing values in reported stress levels were dealt with by means of case-based exclusion. We found a significant main effect for tasks ($F(1, 36) = 38.3$, $p < 0.001$) allowing us to reject the null hypothesis of no difference in reported stress levels.

Table 4 shows the means and standard deviations per task for a selection of features. They indicate, e.g., that the number of SCRs observed in the relaxation tasks is generally lower than in the mentally stressful tasks, as is the average SCL, respiration rate and amplitude. The heart rate variability, measured e.g. through IBI Root Mean Square of Successive Differences (RMSSD), is generally higher in the stressful tasks. Within the mentally stressful or relaxation tasks the three subtasks show similar feature values.

The classification results are shown in Table 5. It can be observed that the performances are very similar over different classification techniques. Comparing the single modalities, the classifiers perform best on ECG, followed by GSR and RSP. Combining features from the three modalities improves performance, and adds an additional 3.5 percentage points on top of the performance obtained using only ECG features. This, however, only when including more than just the set of commonly-used features. For the richest data set covering all features, GMLVQ performs best with just under 87% accuracy. Using 'ordinary' cross validation, performance is slightly higher with accuracies up to 88%.

The diagonal elements of the relevance matrix trained by GMLVQ indicate relevance of individual features to the classifier's decisions. Figure 8 shows these relevances and indicates that most informative features come from the ECG modality, followed by GSR and RSP. The most important individual contributions come from the time domain HRV measures Standard Deviation of Successive Differences (SDSD) and RMSSD, followed by the frequency domain HRV measured by the power in the Very High Frequency (VHF) range. The most influential GSR feature is the minimum SCL and for RSP the mean rate.

5. Discussion

The classification performances for uni-modal stress classification range from 71% for RSP, to 83.4% for ECG indicating that most information can

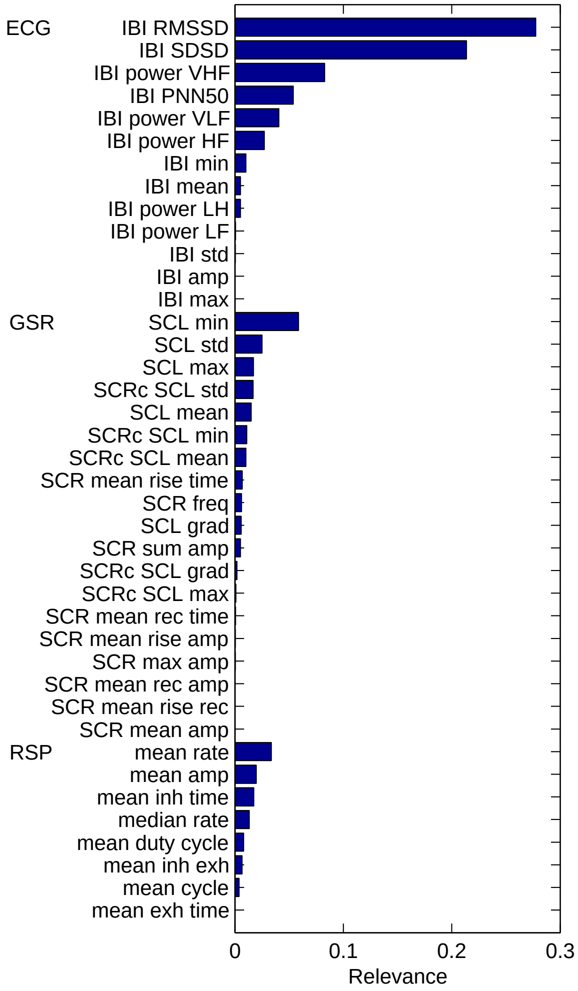

Figure 8: Relevances trained by GMLVQ. For explanation of the feature names used, see Table 3.

be found in the ECG signal. By combining these three modalities the performance is further increased to 86.7% in participant-wise cross-validation. Using 'ordinary' cross-validation we obtained an accuracy of 87.7%. The performances when using two out of the three modalities are better than unimodal performances but lower than the performance when using three modalities and confirm the finding that ECG contains most information to which features from GSR or RSP can provide complimentary information. Both LVQ methods we employed (RSLVQ and GMLVQ) performed well and slightly outperformed the popular SVM, that we included for reference.

With these accuracies, our methods outperform the affective classifiers that are listed in Table 1. They also outperform the participant-wise validated stress-classifiers in Table 2 as well as most others using other cross validation schemes. In comparison to other studies, we used data obtained from a larger number of participants, and also repeated measurements for every participant in 15 sessions spread over several weeks. Thereby, we used a more representative sample of participants and obtained reliable estimates of generalization performance.

The most important feature was found to be the RMSSD. Although it reflects high frequency modulation of heart rate that, in general, is affected by Respiratory Sinus Arrhythmia (RSA), RMSSD has been shown to be unaffected of breathing [56]. The second most influential measure was the related SDSD. The Proportion of IBIs $>$ 50 ms (PNN50) that correlates with RMSSD [45], was also identified as quite relevant. It is worth noting that the two most influential features (RMSSD and SDSD) have been found earlier to be most reliable measures for short term intervals[57], i.e., in the order of 5 minutes, which reflects the measurement time in our experiments. The importance of various HRV measures for distinguishing can be explained by the fact that they reflect parasympathetic (High Frequency (HF) power, RMSSD) and sympathetic (Low Frequency (LF) power, Standard Deviation of IBIs (SDNN)) nervous system responses which are known to relate to the fight-or-flight response and dampening response, respectively [45]. Training SVM and LVQ2.1 after removing irrelevant features, as identified by GMLVQ, did not significantly increase performance. The reduction in noise, however did reduce training times.

Out of the frequency domain HRV measures VHF power we found most influential. Many studies that include cardiac activity as measure for stress only use features from the lower frequency (up to HF) ranges and do not usually consider frequencies above 0.4Hz. This might be due to various reasons. First, the mechanisms that affect VHF power are not well understood. Second, higher frequency ranges cannot be measured reliably through the most commonly used modality BVP which has less sharp peaks, thereby not allowing for a very accurate detection of heart beats which is reflect particularly in inaccuracy in the higher frequencies of HRV. Our use of ECG enabled the reliable use of VHF power as a measure of stress.

We also inspected the prototypes trained by GMLVQ to verify that the stress prototype, as compared to the relax prototype, is characterized by higher heart rate and generally higher HRV values with the exception of RMSSD. Especially for HRV there are varying results published [58, 59, 60], which is confirmed by Berntson and Caciopo, who observed that "it is clear that no single pattern of autonomic adjustments and associated changes in heart rate variability will apply universally across different stressors" [61]. In our study we included three different stressors to induce stressful situations, thereby creating more robustness against this effect. Further, the stress prototype is characterized by more SCRs, higher maximum SCL and faster, though deeper, breathing. These findings are in line with observations made by others [49, 62].

6. Conclusion

We have successfully built classifiers of stress from three physiological modalities and observed that the cardiac activity made the strongest unimodal classifier with over 83% accuracy. Combining the three modalities into a multi-modal classifier improved performance further up to 88% accuracy. By using data from a large sample of participants and repeated sessions we ensured good generalizability to unseen users. The LVQ techniques slightly outperformed well-known techniques such as SVM. These open-box methods allowed us to observe the most important features for stress detection. These were the time domain HRV measures RMSSD and SDSD. The third most important features was found to be very high-frequency HRV from ECG. Most other studies use BVP to measure cardiac activity, however that does not allow for accurate VHF HRV measurements. Therefore it might be advisable for methods that aim at stress detection to use ECG rather than BVP as measurement modality of cardiac activity.

The classifiers built and the knowledge gained on important features for the distinction between stress and relaxation using physiological parameters have brought us one step closer to the realization of a system that can monitor physiology during the day and help its users to monitor their stressful moments during the day. In case a certain quota has been reached or a stressful period reaches a certain duration such a system could trigger the user and offer a means to relief the stress, e.g., a paced breathing exercise [5].

While we have setup our experiments such that they represent daily life as well as possible, the measurements were taken during semi-lab circumstances. Future work should look into the application of the developed classifiers in daily-life measurements and observe their performance. This brings the challenge of reliable ground truth measurements, however this might become more and more feasible with the rapid development of various technologies such as Google Glass [63] that can capture context. It would be interesting to expand the classifier to also be able to classify other affective phenomena such as emotions and moods.

Acknowledgements

The authors would like to thank Geert van Boxtel (Tilburg University), Ad Denissen and Marian Dekker (Philips Research) for providing access to the data set, and Joyce Westerink for her help with the interpretation of the results.

References

[1] E.-M. Backé, A. Seidler, U. Latza, K. Rossnagel, B. Schumann, The role of psychosocial stress at work for the development of cardiovascular diseases: a systematic review, International archives of occupational and environmental health 85 (1) (2012) 67–79.

[2] M. Kivimäki, M. Virtanen, M. Elovainio, A. Kouvonen, A. Väänänen, J. Vahtera, Work stress in the etiology of coronary heart disease–a meta-analysis, Scandinavian journal of work, environment & health 32 (6) (2006) 431–442.

[3] P. J. Rosch, The quandary of job stress compensation, Health & Stress (2001) 1–4.

[4] E. Heber, D. D. Ebert, D. Lehr, S. Nobis, M. Berking, H. Riper, Efficacy and cost-effectiveness of a web-based and mobile stress-management intervention for employees: design of a randomized controlled trial, BMC Public Health 13 (1) (2013) 1.

[5] J. Westerink, W. van Beek, E. Daemen, J. Janssen, G.-J. de Vries, M. Ouwerkerk, The vitality bracelet: Bringing balance to your life with psychophysiological measurements, in: S. H. Fairclough, K. Gilleade (Eds.), Advances in Physiological Computing, Springer London, London, 2014, pp. 197–209.

[6] T. Dempster, D. Vernon, Identifying indices of learning for alpha neurofeedback training, Applied Psychophysiology and Biofeedback 34 (4) (2009) 309–318.

[7] J. H. Gruzelier, A review of the impact of hypnosis, relaxation, guided imagery and individual differences on aspects of immunity and health, Stress 5 (2) (2002) 147–163.

[8] G. J. van Boxtel, A. J. Denissen, M. Jäger, D. Vernon, M. K. Dekker, V. Mihajlović, M. M. Sitskoorn, A novel self-guided approach to alpha activity training, International Journal of Psychophysiology 83 (3) (2012) 282–294. doi:10.1016/j.ijpsycho.2011.11.004.

[9] J. A. Healey, R. W. Picard, Detecting stress during real-world driving tasks using physiological sensors, IEEE Transactions on Intelligent Transportation Systems 6 (2) (2005) 156–166.

[10] J. Zhai, A. Barreto, C. Chin, C. Li, Realization of stress detection using psychophysiological signals for improvement of human-computer interactions, in: IEEE SoutheastCon, 2005. Proceedings, 2005, pp. 415–420.

[11] J. Zhai, A. Barreto, Stress detection in computer users based on digital signal processing of noninvasive physiological variables, Conference proceedings: ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference 1 (2006) 1355–1358.

[12] J. Choi, R. Gutierrez-Osuna, Using heart rate monitors to detect mental stress, in: Sixth International Workshop on Wearable and Implantable Body Sensor Networks, 2009. BSN 2009, 2009, pp. 219–223.

[13] J. Wijsman, B. Grundlehner, H. Liu, H. Hermens, J. Penders, Towards mental stress detection using wearable physiological sensors, in: 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society,EMBC, 2011, pp. 1798–1801.

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

[15] N. Sharma, T. Gedeon, Objective measures, sensors and computational techniques for stress recognition and classification: A survey, Computer Methods and Programs in Biomedicine 108 (3) (2012) 1287–1301.

[16] Neural Networks Research Centre, Helsinki, Bibliography on the self-organizing maps (SOM) and learning vector quantization (LVQ), Otaniemi: Helsinki Univ. of Technology. Available on-line: `http://liinwww.ira.uka.de/bibliography/Neural/SOM.LVQ.html`.

[17] R. S. Lazarus, S. Folkman, Stress, Appraisal, and Coping, Springer Publishing Company, 1984.

[18] R. Sinha, O. A. Parsons, Multivariate response patterning of fear, Cognition and Emotion 10 (2) (1996) 173–198.

[19] R. W. Picard, E. Vyzas, J. Healey, Toward machine emotional intelligence: Analysis of affective physiological state, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (10) (2001) 1175–1191.

[20] K. H. Kim, S. W. Bang, S. R. Kim, Emotion recognition system using short-term monitoring of physiological signals, Medical & Biological Engineering & Computing 42 (3) (2004) 419–427.

[21] C. L. Lisetti, F. Nasoz, Using noninvasive wearable computers to recogniza human emotions from physiological signals, Journal of Applied Signal Processing 11 (2004) 1672–1687.

[22] P. Rani, C. Liu, N. Sarkar, E. Vanman, An empirical study of machine learning techniques for affect recognition in human-robot interaction, Pattern Analysis & Applications 9 (1) (2006) 58–69.

[23] J. Kim, E. André, Emotion recognition based on physiological changes in music listening, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (12) (2008) 2067–2083.

[24] G. Chanel, J. J. M. Kierkels, M. Soleymani, T. Pun, Short-term emotion assessment in a recall paradigm, International Journal of Human-Computer Studies 67 (8) (2009) 607–627.

[25] S. A. Hosseini, M. A. Khalilzadeh, S. Changiz, Emotional stress recognition system for affective computing based on bio-signals, Journal of Biological Systems 18 (1) (2010) 101–114.

[26] E. L. Van den Broek, V. Lisý, J. H. Janssen, J. H. D. M. Westerink, M. H. Schut, K. Tuinenbreijer, Affective man-machine interface: Unveiling human emotions through biosignals, in: A. Fred, J. Filipe, H. Gamboa (Eds.), Biomedical Engineering Systems and Technologies: BIOSTEC2009 Selected Revised papers, Vol. 52 of Communications in Computer and Information Science, Springer, Berlin/Heidelberg, Germany, 2010, pp. 21–47.

[27] C. D. Katsis, N. Katertsidis, G. Ganiatsas, D. I. Fotiadis, Toward emotion recognition in car-racing drivers: A biosignal processing approach, IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans 38 (3) (2008) 502–512.

[28] M. M. Sitskoorn, G. J. M. van Boxtel, J. I. M. Geurdes, D. J. Vernon, A. J. M. Denissen, V. Holten, M. Jaeger, Effectiveness study on a lean-backward audio based neurofeedback training (nft) to enhance cognitive performance, mood and reduce stress of healthy subjects, unpublished study protocol (2009).

[29] G. Logan, W. Cowan, On the ability to inhibit thought and action: A theory of an act of control., Psychological Review 91 (3) (1984) 295–327.

[30] S. Zysset, K. Müller, G. Lohmann, D. von Cramon, Color-word matching stroop task: Separating interference and response conflict, NeuroImage 13 (1) (2001) 29–36.

[31] J. Hammes, De stroop kleur-woord test: handleiding, Swets & Zeitlinger, 1971.

[32] A. Bouma, J. Mulder, L. Lindeboom, Neuropsychologische diagnostiek: Handboek, Swets & Zeitlinger, Lisse, 1996.

[33] W. Kirchner, Age differences in short-term retention of rapidly changing information, Journal of experimental psychology 55 (4) (1958) 352–358.

[34] B. A. Eriksen, C. W. Eriksen, Effects of noise letters upon the identification of a target letter in a nonsearch task, Perception & Psychophysics 16 (1) (1974) 143–149.

[35] S. de Waele, G.-J. de Vries, M. Jager, Experiences with adaptive statistical models for biosignals in daily life, in: Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on, 2009, pp. 1–6.

[36] H. V. Poor, An Introduction to Signal Detection and Estimation, Springer, 1994.

[37] M. Malik, J. T. Bigger, A. J. Camm, R. E. Kleiger, A. Malliani, A. J. Moss, P. J. Schwartz, Heart rate variability standards of measurement, physiological interpretation, and clinical use, European Heart Journal 17 (3) (1996) 354–381.

[38] P. Grossman, E. W. Taylor, Toward understanding respiratory sinus arrhythmia: relations to cardiac vagal tone, evolution and biobehavioral functions, Biological psychology 74 (2) (2007) 263–285. doi:10.1016/j.biopsycho.2005.11.014.

[39] J. J. B. Allen, A. S. Chambers, D. N. Towers, The many metrics of cardiac chronotropy: a pragmatic primer and a brief comparison of metrics, Biological psychology 74 (2) (2007) 243–262.

[40] P. Kohlish, SCRGAUGE - a computer program for the detection and quantification of SCRs, in: W. Boucsein (Ed.), Electrodermal Activity, Plenum, New York, 1992, pp. 432–442.

[41] G.-J. de Vries, M. D. van der Zwaag, Enhanced method for robust mood extraction from skin conductance, in: Proceedings of the third International Conference on Bio-inspired Systems and Signal Processing (BIOSIGNALS), Valencia, Spain, 2010, pp. 139–144.

[42] D. Lemire, Streaming maximum-minimum filter using no more than three comparisons per element, Nordic Journal of Computing 13 (4) (2006) 328–339.

[43] M. E. Dawson, A. M. Schell, D. L. Filion, The electrodermal system, in: J. T. Cacioppo, L. G. Tassinary, G. G. Berntson (Eds.), Handbook

of Psychophsiology, Vol. 2nd, Cambridge University Press, 2000, pp. 200–223.

[44] R. M. Stern, W. J. Ray, K. S. Quigley, Psychophysiological Recording, Oxford University Press, 2001.

[45] Task Force of the European Society of Cardiology the North American Society of Pacing Electrophysiology, Heart rate variability standards of measurement, physiological interpretation, and clinical use, Circulation 93 (5) (1996) 1043–1065. doi:10.1161/01.CIR.93.5.1043.

[46] E. van den Broek, J. Janssen, J. Westerink, J. Healey, Prerequisites for affective signal processing (asp), in: P. Encarnao, A. Veloso (Eds.), Biosignals 2009: Proceedings of the Second International Conference on Bio-Inspired Systems and Signal Processing, INSTICC Press, Portugal, 2009, pp. 426–433.

[47] E. van den Broek, M. van der Zwaag, J. Healey, J. Janssen, J. Westerink, Prerequisites for affective signal processing (asp) - part iv, in: J. Kim, P. Karjalainen (Eds.), Proceedings of the 1st International Workshop on Bio-inspired Human-Machine Interfaces and Healthcare Applications - B-Interface 2010, INSTICC Press, Portugal, 2010, pp. 59–66.

[48] W. Boucsein, Electrodermal activity, Plenum Press, 1992.

[49] W. Boucsein, Electrodermal Activity, Springer, 2012.

[50] T. Kohonen, Improved versions of learning vector quantization, in: International Joint Conference on Neural Networks, Vol. 1, IEEE, 1990, pp. 545–550. doi:10.1109/IJCNN.1990.137622.

[51] S. Seo, K. Obermayer, Soft learning vector quantization, Neural Computation 15 (2003) 1589–1604. doi:10.1162/089976603321891819.

[52] A. W. Witoelar, A. Ghosh, J. J. G. de Vries, B. Hammer, M. Biehl, Window-based example selection in learning vector quantization, Neural Computation 22 (11) (2011) 2924–2961.

[53] P. Schneider, M. Biehl, B. Hammer, Adaptive relevance matrices in learning vector quantization, Neural Computation 21 (12) (2009) 3532–3561. doi:10.1162/neco.2009.11-08-908.

[54] V. Vapnik, Statistical learning theory, Wiley, 1998.

[55] A. Gale, J. A. Edwards, Physiological Correlates of Human Behaviour: Individual differences and psychopathology, Academic Press, 1983.

[56] J. Penttilä, A. Helminen, T. Jartti, T. Kuusela, H. V. Huikuri, M. P. Tulppo, R. Coffeng, H. Scheinin, Time domain, geometrical and frequency domain analysis of cardiac vagal outflow: effects of various respiratory patterns, Clinical Physiology 21 (3) (2001) 365376.

[57] J. McNames, M. Aboy, Reliability and accuracy of heart rate variability metrics versus ECG segment duration, Medical and Biological Engineering and Computing 44 (9) (2006) 747–756.

[58] K. J. Mathewson, M. K. Jetha, I. E. Drmic, S. E. Bryson, J. O. Goldberg, G. B. Hall, D. L. Santesso, S. J. Segalowitz, L. A. Schmidt, Autonomic predictors of stroop performance in young and middle-aged adults, International Journal of Psychophysiology 76 (3) (2010) 123–129.

[59] C. E. Wright, K. O'Donnell, L. Brydon, J. Wardle, A. Steptoe, Family history of cardiovascular disease is associated with cardiovascular responses to stress in healthy young men and women, International Journal of Psychophysiology 63 (3) (2007) 275–282.

[60] V. Vuksanović, V. Gal, Heart rate variability in mental stress aloud, Medical Engineering & Physics 29 (3) (2007) 344–349.

[61] G. G. Berntson, J. T. Cacioppo, Heart rate variability: Stress and psychiatric conditions, in: M. Malik, A. J. Camm (Eds.), Dynamic Electrocardiography, Blackwell Publishing, 2004, pp. 57–64.

[62] J. H. Houtveen, S. Rietveld, E. J. de Geus, Contribution of tonic vagal modulation of heart rate, central respiratory drive, respiratory depth, and respiratory frequency to respiratory sinus arrhythmia during mental stress and physical exercise, Psychophysiology 39 (4) (2002) 427–436.

[63] Google, Google glass.
URL `http://www.google.com/glass/start/`

J.J.G. (Gert-Jan) de Vries received the B.Sc. and M.Sc. (cum laude) degrees in computing science from the Groningen University, The Netherlands, in 2005 and 2007, respectively. He has been working as research scientist with Philips Research, Eindhoven, The Netherlands, since 2007. He is also working towards the Ph.D. degree in the Johann Bernoulli Institute for Mathematics and Computer Science, Groningen University, The Netherlands. His work focusses on data analysis, machine learning and intelligent algorithm development applied to topics ranging from emotions, and physiological signals, to healthcare information.

Steffen C. Pauws holds a master in Medical Informatics from the University of Leiden, two post-master degrees in 'Software Technology' and 'Perceptual & Cognitive Engineering' and a PhD degree from the Eindhoven Technical University. He is a member of IEEE. Since 2000, he works for Philips Research being involved and in the lead of algorithms, data analysis and user experiences. His current interests are in the infrastructure and analysis of healthcare data for researching home healthcare services, clinical algorithms and stratification principles for patients with long-term medical conditions.

Michael Biehl received a Ph.D. in Physics from the University of Gießen, Germany, in 1992 and completed a Habilitation in Theoretical Physics at the University of Würzburg, Germany in 1996. He is currently Professor of Computer Science at the Johann Bernoulli Institute for Mathematics and Computer Science, University of Groningen, The Netherlands. His main research interest is in the theory of machine learning techniques and their application in the life sciences. He is furthermore active in the modelling and simulation of complex physical system. He has co-authored more than 150 publications in international journals and conference proceedings.

Table 1: Review of ten machine learning studies employing different physiological signals to recognize various affective states.

Reference	Modalities[1]	Ss[2]	Feat.[3]	Technique	Targets	Perf[4]
Sinha & Parsons (1996)	M	27	18	LDA	2 emotions	86%
Picard et al. (2001)	C,E,R,M	1	40	LDA	8 emotions	81%
Kim et al. (2004)	C,E,S	175		SVM	3 emotions	73%
Lisetti & Nasoz (2004)	C,E,S	29		kNN, LDA, ANN	6 emotions	86%
Rani et al. (2006)	C,E,S,M,P	15	46	kNN, SVM, RT, BN	3 emotions	86%
Kim & André (2008)	C,E,M,R	3	110	LDA, EMDC[5]	4 emotions	79%
Chanel et al. (2009)	C,E,R	11	18		3 emotions	66%
	B		18720		3 emotions	73%
	B,C,E,R		18738		3 emotions	70%
Hosseini et al. (2010)	C,E,R	15	38	SVM	2 arousal levels	77%
	B	15	21	LDA, SVM	2 arousal levels	85%
Van den Broek et al. (2010)	E,M	21	10	kNN, SVM, ANN	4 emotions	61%
Katsis et al. (2008)	C,E,M,R	10	15	SVM, ANFIS	4 affect states	79%

[1] Abbreviations used: B Brain activity (EEG); C Cardiovascular activity (e.g., ECG and BVP);E Electrodermal activity (EDA); M Electromyogram (EMG); P Blood pressure; R Respiration; S Skin temperature

[2] Number of subjects

[3] Number of features

[4] Performance (accuracy)

[5] A tailored ensemble of binary classifiers

Table 2: Review of machine learning studies employing different physiological signals to recognize stress.

Reference	Modalities[1]	Ss[2]	Technique	Targets	Val.[3]	Perf[4]
Healey & Picard (2005)	C,E,R,M	9	LDA	3-level	CV	97%
Zhai et al. (2005)	E,C,O	6	SVM (linear kernel)	2-class	CV	57%
			SVM (RBF kernel)			60%
			SVM (sigmoid kernel)			80%
Zhai & Barreto (2006)	E,C,O,S	32	SVM	2-class	CV	90%
	C,O,S					90%
	E,O,S					90%
	E,C,S					61%
	E,C,O					89%
Choi & Gutierrez-Osuna (2009)	C,R	3	unspecified	2-class	CV within pp	83%
					pp-wise CV	69%
Wijsman et al. (2011)	C,R,E,M	21	Linear Bayes Normal	2-class	CV	78%
			Quadratic Bayes Normal			78%
			kNN			76%
			Fisher's Least Square			79%
Giakoumis et al. (2013)	E	24	LDA	2-class	CV	83%
	C					74%
	E,C					95%
	E,C				pp-wise CV	86%

[1] Abbreviations used: B Brain activity (EEG); C Cardiovascular activity (e.g., ECG and BVP);E Electrodermal activity (EDA); M Electromyogram (EMG); O Ocular Response (e.g., Pupil diameter); P Blood pressure; R Respiration; S Skin temperature

[2] Number of subjects

[3] Type of validation

[4] Performance (accuracy)

Table 3: Features extracted from the raw and preprocessed signals.

ECG	IBI min	minimal IBI
	IBI max	maximal IBI
	IBI mean [*]	mean IBI
	IBI std [*]	standard deviation of IBIs, also referred to as SDNN
	IBI amp	amplitude of IBIs (max-min)
	IBI power VLF [*]	power of IBIs in very low frequency band ($0 - 0.04$ Hz)
	IBI power LF [*]	power of IBIs in low frequency band ($0.04 - 0.15$ Hz)
	IBI power HF [*]	power of IBIs in high frequency band ($0.15 - 0.4$ Hz)
	IBI power VHF	power of IBIs in very high frequency band ($0.4 - 1$ Hz)
	IBI power LH [*]	ratio between IBI power LF and HF
	IBI RMSSD [*]	root mean square of successive differences of IBIs
	IBI PNN50	proportion of IBIs > 50 ms
	IBI SDSD	standard deviation of successive differences of IBIs
GSR	SCL mean [*]	mean SCL
	SCL std [*]	standard deviation of SCL
	SCL grad	gradient of SCL (estimated by best linear fit)
	SCL min	minimal SCL
	SCL max	maximal SCL
	SCR freq [*]	number of SCRs per second
	SCR max amp	maximal amplitude of SCRs
	SCR mean amp [*]	mean amplitude of SCRs
	SCR sum amp	sum of amplitudes of SCRs
	SCR mean rise time [*]	mean rise time of SCRs
	SCR mean rec time [*]	mean half recovery time of SCRs
	SCR mean rise rec	mean ratio of rise time and half recovery time of SCRs
	SCR mean rise amp	mean ratio of rise time and amplitude of SCRs
	SCR mean rec amp	mean ratio of half recovery time and amplitude of SCRs
	SCRc SCL mean	mean SCL after correcting for SCRs
	SCRc SCL std	standard deviation of SCL after correcting for SCRs
	SCRc SCL grad	gradient of SCL after correcting for SCRs (estimated by best linear fit
	SCRc SCL min	minimal SCL after correcting for SCRs
	SCRc SCL max	maximal SCL after correcting for SCRs
RSP	mean rate [*]	mean respiration rate
	median rate	median respiration rate
	mean amp [*]	mean amplitude of respirations
	mean inh time	mean inhalation time
	mean exh time	mean exhalation time
	mean cycle	mean respiration time
	mean duty cycle	ratio between mean inhalation time and cycle
	mean inh exh	ratio between mean inhalation and exhalation time

[*] This feature is included in the commonly used set of features representing all modalities.

Table 4: Statistics (means and standard deviations) per feature per task of non-normalized data. See Table 3 for explanation of the feature names used.

Feature name | Task	Relax 1	Relax 2	Relax 3	N-back	Stop-signal	Stroop
SCR freq	0.029 ± 0.034	0.029 ± 0.036	0.031 ± 0.036	0.057 ± 0.057	0.059 ± 0.059	0.062 ± 0.063
SCL mean	2.043 ± 1.398	2.122 ± 1.454	2.184 ± 1.460	2.447 ± 1.537	2.304 ± 1.518	2.416 ± 1.554
SCL std	0.248 ± 0.201	0.244 ± 0.197	0.242 ± 0.193	0.215 ± 0.184	0.240 ± 0.209	0.237 ± 0.195
RSP rate	15.432 ± 3.827	15.080 ± 3.867	14.951 ± 3.877	16.303 ± 3.700	16.488 ± 3.391	16.399 ± 3.580
RSP mean amp	6.367 ± 4.755	6.452 ± 5.016	6.419 ± 4.978	6.947 ± 5.124	6.781 ± 5.114	7.048 ± 5.766
IBI mean	0.825 ± 0.112	0.839 ± 0.113	0.853 ± 0.116	0.842 ± 0.115	0.835 ± 0.113	0.836 ± 0.114
IBI RMSSD	0.053 ± 0.034	0.057 ± 0.039	0.060 ± 0.038	0.073 ± 0.056	0.071 ± 0.052	0.071 ± 0.050

Table 5: Generalization performance for the three classifiers on the five feature sets for both cross validation schemes. The numbers listed are means and standard deviations over 10x10-folds of validation

pp-wise cross validation			
	SVM	RSLVQ	GMLVQ
RSP	71.0% ± 0.2%	70.9% ± 0.3%	71.0% ± 0.3%
GSR	74.8% ± 0.3%	74.8% ± 0.2%	74.4% ± 0.3%
ECG	83.2% ± 0.1%	83.1% ± 0.1%	83.4% ± 0.2%
RSP&GSR	77.5% ± 0.2%	77.6% ± 0.2%	77.2% ± 0.2%
RSP&ECG	85.2% ± 0.2%	84.8% ± 0.2%	85.4% ± 0.2%
GSR&ECG	85.5% ± 0.1%	85.2% ± 0.2%	85.6% ± 0.2%
RSP&GSR&ECG-selection	81.4% ± 0.3%	81.6% ± 0.3%	81.6% ± 0.2%
RSP&GSR&ECG-all	86.6% ± 0.2%	86.6% ± 0.2%	86.7% ± 0.2%
cross validation			
	SVM	RSLVQ	GMLVQ
RSP	71.8% ± 0.1%	71.6% ± 0.1%	71.8% ± 0.1%
GSR	75.3% ± 0.1%	75.4% ± 0.1%	75.0% ± 0.2%
ECG	83.6% ± 0.1%	83.5% ± 0.1%	83.6% ± 0.1%
RSP&GSR	78.4% ± 0.2%	78.5% ± 0.2%	78.2% ± 0.1%
RSP&ECG	86.0% ± 0.1%	85.8% ± 0.1%	86.2% ± 0.2%
GSR&ECG	86.0% ± 0.1%	85.8% ± 0.1%	86.2% ± 0.1%
RSP&GSR&ECG-selection	82.7% ± 0.1%	82.5% ± 0.1%	82.6% ± 0.1%
RSP&GSR&ECG-all	87.6% ± 0.1%	87.7% ± 0.2%	87.6% ± 0.1%